Master's Thesis

Systematic Duration Allocation in Government Bond Indices

Edoardo Tarcisio Visconti

Supervisors: David Anderson (OLZ), Dr. Patrick Walker (OLZ), Dr. Urban Ulrych (ETH), Prof. Patrick Cheridito (ETH)

Department of Mathematics

Fall Semester 2025

Abstract

This thesis studies systematic duration-timing strategies in government bond markets, testing whether dynamic maturity allocation delivers risk-adjusted outperformance relative to an equal-weighted benchmark. The analysis is conducted using government bond indices segmented into four maturity buckets over the period 2005–2024. Statistical inference is conducted using HAC-based tests and block bootstrap procedures applied to net Sharpe ratio differences.

The empirical investigation proceeds in several steps. First, factor-based strategies are examined on the World Government Bond Index (WGBI). Carry, value, and momentum signals contain economically meaningful information, but their standalone performance is limited and regime dependent. Carry, as well as combinations that weight signals according to their past cross-sectional correlation with returns (the information coefficient, IC, approach), deliver the most robust results, with economically meaningful gains and partial statistical support, and represent the most defensible choices for practitioners in this setting.

Second, the focus shifts to U.S. Treasuries. Applying the same strategies, we find that outperforming the equal-weighted benchmark in terms of Sharpe ratios is substantially more difficult, and statistical significance is generally weak. The information set is therefore extended to include macroeconomic, market, and exchange-rate variables. Duration timing is implemented using machine-learning methods under strict constraints on model complexity and feature selection. While specific subsets of predictors generate out-of-sample gains that remain economically meaningful after transaction costs, statistical evidence of outperformance relative to the benchmark remains limited.

Third, simple threshold-based strategies are introduced to capture stress regimes using policy, inflation, volatility, and yield-curve indicators. These rule-based approaches underperform in stable environments but provide effective protection during periods of market stress, although they do not exhibit strong statistical significance.

Finally, the analysis is extended to a multi-country setting using Instrumented Principal Component Analysis (IPCA). Mapping IPCA forecasts into duration-based portfolio allocations across countries and maturities yields out-of-sample improvements and overall encouraging results. Under selected specifications, statistical significance at the 5% level is achieved, while alternative configurations remain economically meaningful but less statistically robust.

Overall, the results indicate that active duration timing can generate economic value when signals are carefully combined and model complexity is tightly controlled. However, achieving statistical significance remains challenging, particularly in the U.S. Treasury market, likely reflecting its relatively high informational efficiency. From a practical perspective, several of the proposed strategies remain implementable and economically viable, offering structured tools for practitioners engaged in duration management.

Contents

1 Introduction and Literature

1.1 Problem Setting and Motivation

Fixed-income markets play an important role in the financial system, particularly for institutional investors. Government bonds are widely used to manage risk, construct portfolios, and serve as benchmarks. Although they are often perceived as safe assets, bond portfolios are actively managed in practice, mainly to control exposure to interest-rate risk and to achieve stable risk-adjusted returns. A central decision in this context is how to allocate capital across bonds with different duration, balancing interest-rate sensitivity, return potential, and downside risk as market conditions evolve over time.

This decision is known as duration allocation or duration timing. Bonds with different duration exposures react differently to changes in interest rates, inflation, and broader market conditions. Portfolios with higher duration are more sensitive to interest-rate movements and tend to perform well when growth is stable and inflation expectations are anchored. Portfolios with lower duration, in contrast, provide greater protection when interest rates rise or uncertainty increases. Because economic conditions and market environments evolve over time, the relative attractiveness of different duration exposures also changes. For this reason, maintaining a fixed duration exposure can be suboptimal, motivating dynamic approaches to duration allocation. This issue is further amplified by the strong regime dependence of bond returns and by the limited scope for diversification within government bond portfolios, where returns across duration exposures are highly correlated.

In this thesis, we address the problem of duration allocation in a systematic and empirical manner with a twofold objective. First, we conduct an empirical analysis, documenting how relative performance across maturities evolves over time and how it relates to observable signals. Second, we design and evaluate allocation strategies that are directly usable by asset managers and finance professionals who regularly deal with bond investments.

Much of the existing academic literature on bond return predictability (including the work of Cieslak and Povala [7], Cochrane and Piazzesi [8], Diebold and Li [9], and Ludvigson and Ng [19]), focuses on forecasting bond excess returns across maturities and countries, typically treating duration as a continuous characteristic or implicitly aggregating information across the yield curve. In contrast, this thesis addresses a different and more comparative question. Rather than predicting bond returns in levels, we ask which maturity segment offers the most attractive risk-adjusted opportunity at a given point in time for a specific country, or for a set of countries in the multi-country setting. Framed this way, the problem is closer to a ranking or classification task across duration buckets, with the objective of identifying the most favorable maturity to hold at each rebalancing date. This thesis also fits into the ongoing debate on the predictability of bond returns by taking a portfolio perspective. Rather than focusing solely on statistical predictability from signals, we explicitly examine whether such signals can be translated into implementable portfolio strategies. In this sense, our analysis is closer to the economic evaluation emphasized by Thornton and Valente [20].

To this end, we focus on government bond indices rather than individual securities. These indices provide liquid and investable representations of sovereign bond markets and are naturally organized by maturity, which serves as a proxy for duration exposure in our setting. Duration timing is implemented by allocating capital across four maturity buckets (1–3, 3–5, 5–10, and 10+ years) of the selected government bond index or indices. At each rebalancing date, the goal is to shift exposure across these buckets in order to identify which part of the yield curve offers the highest risk-adjusted return over the subsequent investment period. Strategy performance is evaluated primarily using out-of-sample Sharpe ratios, reflecting the perspective of a professional asset manager focused on stable risk-adjusted performance rather than in-sample fit. Cumulative returns are reported for completeness, but play a secondary role in the assessment.

All strategies analyzed here use a holding period of 21 business days, corresponding to a monthly investment horizon commonly used in practice. The analysis is restricted to long-only allocations, in line with typical institutional constraints that rule out short positions in sovereign bond segments. Cash positions are not considered, as cash-like instruments behave similarly to very short-maturity government bonds and therefore do not represent a separate allocation choice.

The analysis combines market data with selected macroeconomic series from the Federal Reserve Economic Data (FRED) database. Although data are available from 2000 to the end of 2024, the sample starts in 2005 to ensure enough history for reliable signal construction. The period from 2005 to 2013 is used to train the models, with the aim of improving robustness rather than maximizing in-sample performance. Results are then evaluated strictly out of sample over the period 2014–2024.

Because bond markets behave very differently across economic regimes, these choices are meant to favor strategies that work reasonably well in different environments rather than fitting a specific historical period. To highlight these differences, performance is also reported separately for the subperiods 2014–2019 and 2019–2024. This split is used only for evaluation and does not affect model design or calibration.

Alongside the standard Sharpe ratio, we also report a net Sharpe ratio that accounts for transaction costs. Returns are adjusted by subtracting a linear cost proportional to portfolio turnover,

$$r_t^{\text{net}} = r_t - c \cdot \text{turnover}_t,$$

and the net Sharpe ratio is then computed using the resulting series r_t^{net}. We set the cost parameter to $c = 5$ bps per unit of turnover for WGBI and U.S.-only portfolios, and to $c = 10$ bps for the multi-country portfolio to reflect the lower liquidity of some constituent markets.

Turnover is computed as the sum of absolute changes in portfolio weights between the pre-rebalance allocation and the target allocation,

$$\text{turnover}_t = \sum_i \left| w_{i,t}^{\text{target}} - w_{i,t}^{\text{pre}} \right|.$$

In this work, in specific selected sections below, statistical significance is formally assessed through p-values associated with net Sharpe ratio comparisons. These are

computed using the robust testing framework of Ledoit and Wolf [17], which accounts for heteroskedasticity and serial correlation in returns (HAC), building on the asymptotic results of Lo [18]. Specifically, we conduct relative Sharpe ratio tests that evaluate whether the difference in risk-adjusted performance (after transaction costs) between a given strategy and a benchmark is statistically greater than zero.

The benchmark used throughout is an equal-weighted portfolio constructed over the same investment universe and evaluation window as the strategy under consideration. The precise definition of the equal-weighted portfolio is specified in each relevant section.

The implementation relies on a modified version of the open-source library RobustSharpeRatioHAC (GitHub), which closely follows [17] and [18]. The original code performs two-sided tests; it has been adapted here to conduct one-sided tests, as the object of interest is strictly outperformance relative to the benchmark. Both asymptotic HAC-based inference and a block bootstrap procedure are employed.

Given the relatively limited effective sample size, approximately 120 monthly rebalancing decisions over the full evaluation period (2014–2024), a block bootstrap procedure is employed alongside asymptotic HAC-based inference. The bootstrap uses circular blocks of six months in order to preserve the serial dependence structure of returns while providing an alternative approximation of the sampling distribution of the Sharpe ratio difference. It is reported primarily as a robustness check, allowing us to verify whether the conclusions obtained under HAC inference remain coherent in a limited-sample setting and are not driven by small-sample distortions. Given this time-series dimension, the statistical power of the tests is not expected to be high, so moderate p-values should be interpreted with caution rather than as evidence against economically meaningful differences.

We choose not to apply a multiple-testing correction in this setting. The analysis does not rely on a large-scale search over randomly generated strategies, but rather on a limited set of economically motivated and theoretically grounded specifications. For this reason, we focus on the individual p-values associated with each comparison. The implications of multiple testing are nonetheless discussed separately, in order to clarify how the reported significance levels should be interpreted within the broader empirical framework.

1.2 Methodology, Literature, and Findings

The empirical analysis in this thesis is organized into three steps, each building on the previous one and gradually increasing the dimensionality and scope of the problem. Across all parts, the common objective is to study duration timing, while progressively enriching the information set and the modeling framework, from factor strategies to more flexible models.

a. Factor Strategies on the WGBI Index

The first part of the thesis focuses on the World Government Bond Index (WGBI). It is a global benchmark that aggregates sovereign bond markets across developed countries and maturity segments (with country weights determined by the outstanding market

value of eligible government bonds). By construction, the WGBI is segmented by maturity, making it a suitable setting for studying duration allocation.

Within this framework, we examine whether well-established style factors (such as Value, Momentum, and Carry) contain predictive information for duration timing. In the context of government bonds, these factors summarize simple and economically motivated signals derived from prices and yields: Value captures whether yields are high or low relative to their historical level, Momentum reflects the persistence of recent price or yield movements, and Carry measures the income earned from holding bonds under the assumption that the yield curve remains unchanged.

This analysis is closely related to the evidence in Brooks and Moskowitz [5] and Baltussen, Martens, and Penninga [2], who show that such style factors explain a substantial share of yield-curve premia across countries and maturities, often outperforming traditional term-structure factors such as level, slope, and curvature. Related work by Asness, Moskowitz, and Pedersen [1] documents that value and momentum effects are persistent across asset classes, including government bonds, while the broader trend-following literature (including Hurst, Ooi, and Pedersen [11], and Kolanovic and Wei [16]) shows that price trends are a general feature of asset returns. Building on this literature, we do not restrict attention to individual factor signals alone, but also study combinations of signals and variants that incorporate additional features or conditioning information, with the aim of assessing whether richer factor constructions improve duration-timing decisions.

Taken together, the empirical evidence indicates that factor signals contain economically meaningful information for duration timing, yet their standalone performance is generally fragile. With the exception of Carry, individual factors fail to deliver strong and stable risk-adjusted returns across regimes, suggesting that reliance on a single signal is rarely sufficient.

From a practical allocation perspective, and in light of the formal Sharpe ratio tests, the evidence supports either Carry-based specifications, given their statistically significant outperformance, or an allocation rule grounded in information coefficients (ICs). The IC strategy weights maturities proportionally to the estimated cross-sectional correlation between each signal and subsequent returns. In practice, the predictive strength of Carry, Value, and Momentum is assessed through their historical correlation with returns, and capital is allocated accordingly: signals with stronger empirical correlation receive larger weights, while weaker signals are scaled down.

Although this IC-based allocation falls marginally short of conventional significance thresholds, it relies on a transparent statistical principle and exhibits comparatively robust empirical performance, making it defensible from both an economic and methodological standpoint.

b. Strategies for U.S. Bond Markets Using Macroeconomic and Market Signals with Machine Learning

The second part of the thesis shifts attention to a single-country setting, focusing on U.S. government bonds. We first apply the same factor strategies used in the global WGBI analysis. In this setting, their performance is generally weaker and less stable, and, crucially, we find no statistical evidence of outperformance relative to the bench-

mark. A natural interpretation is that the U.S. yield curve exhibits a higher degree of informational efficiency, leaving limited scope for term structure–based signals to generate economically and statistically meaningful duration-timing gains. This motivates broadening the analysis beyond factor signals derived solely from the yield curve.

At the country level, it becomes possible to incorporate a richer set of macroeconomic and market variables that are inherently country specific, such as policy rates, inflation measures, indicators of economic activity, and financial market conditions. A large literature documents that these variables contain predictive information for bond excess returns and that their effects differ across maturities. Early evidence by Ilmanen [12] shows that simple economic and market indicators forecast returns at different points of the yield curve in distinct ways, directly supporting the idea of rotating across maturity buckets. Ludvigson and Ng [19] extend this framework by extracting latent macroeconomic factors from large information sets, demonstrating that macro variables improve bond return forecasts beyond yield-curve information alone.

A practical challenge in this setting is how to integrate these new predictors into implementable allocation strategies. To address this issue, we rely on machine learning methods, which allow for flexible modeling of non-linearities and interactions. This approach is motivated by recent evidence showing that bond return predictability is strongly regime dependent and benefits from data-driven models. In particular, Bianchi, Buchner, and Tamoni [4] show that machine learning techniques extract additional predictive structure in U.S. Treasury markets, especially when dealing with a large set of macroeconomic and financial variables, and that predictive patterns vary across maturities. Complementary evidence by Caruso and Coroneo [6] highlights the importance of real-time macroeconomic information, showing that predictive content depends on the information set available to investors at the time of the decision.

In our empirical results, machine learning approaches provide partial improvements when combined with appropriate information sets, particularly Forex-based variables and selected macroeconomic indicators. However, their performance is strongly regime dependent and varies across market conditions. Simpler models tend to generalize more reliably out of sample, while more complex specifications are more prone to overfitting and deliver less stable performance. From a statistical standpoint, the evidence is weak: p-values are generally far from conventional significance thresholds.

In addition to machine learning approaches, we also test simple threshold-based strategies. These rule-based methods prioritize robustness and interpretability, limiting overfitting in a data-constrained environment. While they tend to underperform in stable, falling-yield regimes, they are most effective during periods of market stress, serving as a low-complexity and defensive benchmark for duration allocation.

We also consider combinations of the different strategy classes discussed above by forming ensemble allocations across factor based, machine learning, and threshold strategies. The results suggest that such ensembles add limited incremental value, as overall performance is largely driven by a small number of components rather than by diversification across strategy types.

Taken together, these results indicate that the U.S. Treasury market represents a particularly demanding environment for active duration timing. Across all approaches, beating the equal-weighted benchmark in a consistent and statistically robust manner proves difficult. While some strategies are economically sensible and few exhibit

episodes of good performance, none achieves statistically significant outperformance at conventional levels once benchmark exposure and sampling uncertainty are properly accounted for.

c. Strategies Using IPCA Across Multiple Bond Indices and Countries

The third part extends the analysis to a multi-country, multi-index setting using Instrumented Principal Component Analysis (IPCA). Introduced by Kelly, Pruitt, and Su [14], IPCA is a conditional factor model in which latent risk factors are estimated jointly with time-varying factor loadings that depend on observable characteristics. This framework connects traditional latent factor models with characteristic-based approaches, making it possible to identify common sources of risk while allowing exposures to differ across countries, maturities, and over time. As shown in subsequent work [15], characteristics should be interpreted as indicators of how assets' risk exposures change with underlying factors, rather than as variables that directly predict returns. More broadly, the IPCA procedure can be viewed as a linear analogue of an autoencoder, as discussed by GU, Kelly, and Xiu in [10].

The empirical usefulness of IPCA has been demonstrated in both equity and corporate bond markets. In particular, Kelly, Palhares, and Pruitt [13] show that IPCA substantially improves the explanation of realized variation and expected returns in corporate bond data. Building on this evidence, we apply IPCA to government bond indices and duration timing. In our setting, IPCA is used to forecast expected returns across countries and maturity buckets, and these forecasts are then translated into portfolio decisions either by ranking maturities or by constructing mean–variance portfolios. Both approaches deliver out-of-sample performance improvements relative to an equal-weighted benchmark, confirming that IPCA provides a flexible framework for modeling and exploiting time-varying expected returns along the yield curve in a multi-country context. Under selected specifications (using categorical variables for duration buckets and a softmax allocation rule), statistical significance at the 5% level is achieved, while alternative configurations remain economically strong but less statistically robust.

Importantly, the results show that these strategies successfully adjust duration exposure over time, effectively acting as dynamic duration selectors rather than relying primarily on cross-country allocation.

Overall, the evidence in this section is encouraging. The structure of the IPCA-based allocations is straightforward and implementable, making the framework relevant not only from an academic standpoint but also for practitioners engaged in duration management.

2 Government Bonds and the Macroeconomic Framework

In this introductory section, we review the fundamental mechanics of government bonds, formalizing the relationship between prices and yields and examining how both respond to monetary policy actions and broader macroeconomic conditions. We then introduce the concept of duration as a measure of interest rate sensitivity and discuss the economic interpretation and typical shapes of the yield curve. This framework motivates the focus on maturity-specific portfolios and duration-timing strategies developed in the remainder of the thesis.

2.1 Government Bonds

A bond is a debt instrument in which the issuer promises to pay fixed coupon interest at regular intervals and to repay the principal at maturity. Government bonds (or sovereign bonds) are issued by national governments to borrow from the market and finance public spending.

At issuance, a bond has:

- a redemption value R (the amount repaid at maturity);
- a sequence of cash flows $\{C_t\}_t$, usually a fixed percentage of the face value (the coupon).

The price of the bond equals the present value of future payments discounted at the market yield to maturity (YTM).

$$P = \sum_{t=1}^{T} \frac{C_t}{(1 + \text{YTM})^t} + \frac{R}{(1 + \text{YTM})^T}.$$

The YTM is the bond's internal rate of return if held to maturity: the single discount rate that makes the present value of all coupons and principal equal to the market price. Plotting yields against maturities gives the *yield curve*. In this project we use bond indices, which aggregate bonds across maturities. Each index's yield is the market-value-weighted average of its constituents' yields.

In the *primary market* (issuance stage), the yield emerges from investor demand; the price adjusts so that the fixed coupon rate and market yield are consistent. In the *secondary market*, yields continue to fluctuate dynamically as bonds are traded and market conditions evolve.

From the equation above, price and yield move inversely:

$$\text{Yield rises} \Rightarrow \text{Price falls}, \qquad \text{Yield falls} \Rightarrow \text{Price rises}.$$

Market prices adjust so that a bond's yield equals the required return for its maturity and risk profile. In equilibrium, bonds with identical risk and maturity characteristics must offer the same yield, ensuring the absence of arbitrage.

When a new bond is issued, it establishes a benchmark yield for that maturity. Existing bonds with similar maturities reprice to align with that benchmark, including any liquidity premia, and local yield changes propagate through the entire yield curve.

A bond's sensitivity to changes in yield is measured by its duration. The *Macaulay duration* D_{mac} is defined as the weighted average time of the discounted cash flows:

$$D_{\text{mac}} = \frac{1}{P} \sum_{t=1}^{T} t \cdot \frac{C_t + I_{t=T} R}{(1 + \text{YTM})^t}.$$

It measures the average time (in years) required to recover the bond's price through its discounted payments.

The *modified duration* adjusts for compounding:

$$D_{\text{mod}} = \frac{D_{\text{mac}}}{1 + \text{YTM}}.$$

For a small change in yield Δy,

$$\frac{\Delta P}{P} \approx -D_{\text{mod}} \, \Delta y.$$

Thus, bonds with longer duration are more sensitive to yield movements. Intuitively, the Macaulay duration represents the time at which all discounted cash flows could be concentrated into a single equivalent payment.

The *convexity* refines this linear approximation by capturing curvature in the price–yield relation:

$$\frac{\Delta P}{P} \approx -D_{\text{mod}} \, \Delta y + \tfrac{1}{2} C_x (\Delta y)^2,$$

where C_x denotes the convexity coefficient. Higher convexity implies that price gains from yield decreases are larger than price losses from equivalent yield increases.

This framework provides the mathematical foundation to interpret bond prices, yields, and their sensitivity to macroeconomic and policy shifts.

2.2 Yield Curve and Monetary Policy

a. Yield Curve Dynamics

The yield curve, i.e., the term structure of interest rates, describes how bond yields vary with their maturities at a given point in time.

The (zero-coupon) yield curve is widely employed as a benchmark for pricing fixed-income securities and as an indicator of economic expectations. Its configuration provides insights into anticipated economic growth, inflation, and monetary policy direction:

- A **normal** (upward-sloping) curve, where longer maturities exhibit higher yields, indicates a positive term premium. This is the most common configuration. Longer-term bonds incorporate both expectations of future interest rate increases in an expanding economy and a higher yield to compensate investors for greater liquidity risk over time.

- An **inverted** (downward-sloping) curve, where short-term yields exceed long-term yields, reflects expectations that current restrictive monetary policy will lead to slower growth and eventual rate cuts. It typically signals market anticipation of an economic downturn or recession, as observed before major crises such as 2008.

The curve can also be **flat**, reflecting uncertainty or a transitional phase in the economic cycle, where short- and long-term yields converge due to ambiguous expectations about future interest rates. Alternatively, a **humped** curve, with yields peaking at intermediate maturities, may indicate temporary market imbalances in supply and demand or differing risk perceptions across maturities.

Table 1: Yield Curve Shapes and Historical Examples

Yield Curve Shape	**Examples and Context**
Normal (upward-sloping)	U.S. (2004–2006): typical of an expansion phase. Short-term rates were low while long-term yields were higher, reflecting expectations of future growth, inflation, and monetary tightening.
Inverted (downward-sloping)	U.S. (2006–2007): short-term yields exceeded long-term yields as the Federal Reserve raised policy rates. The inversion anticipated the 2008 financial crisis.

The Nelson–Siegel model provides a parsimonious parametric representation of the yield curve. It expresses the yield at maturity τ as:

$$y(\tau) = \beta_0 + \beta_1 \frac{1 - e^{-\tau/\lambda}}{\tau/\lambda} + \beta_2 \left(\frac{1 - e^{-\tau/\lambda}}{\tau/\lambda} - e^{-\tau/\lambda} \right)$$

The three parameters β_0, β_1, and β_2 (obtained in practice by minimising the MSE at each time) correspond respectively to the *level*, *slope*, and *curvature* of the yield curve, while λ controls the exponential decay rate that determines the maturity at which the maximum curvature occurs. These three factors (level, slope, and curvature) are also those typically identified by applying PCA to yields across maturities, confirming that both methods capture the same fundamental sources of variation in the term structure.

A more flexible extension of this specification is the Nelson–Siegel–Svensson (NSS) model, which adds an additional curvature term governed by a second decay parameter. This introduces an extra coefficient that allows the yield curve to capture more complex shapes, particularly at long maturities. In the empirical analysis that follows, we adopt this extended specification to better fit the cross section of yields while retaining a clear economic interpretation of the underlying factors.

b. Monetary Policy

Central banks use policy rates to steer inflation, employment, and overall economic stability. By influencing short-term borrowing costs, policy rates affect aggregate demand, credit conditions, and inflation expectations, with the ultimate goal of keeping inflation close to target while stabilizing output.

Rate hikes are implemented to slow economic activity and contain inflation. Higher borrowing costs reduce consumption and investment, short-term yields rise, and the yield curve often flattens or inverts as markets anticipate weaker growth or future easing. Rate cuts, instead, aim to stimulate the economy: borrowing becomes cheaper, demand recovers, short-term yields fall, and the curve typically steepens. When policy is stable and inflation is near target, yield movements are driven mainly by expectations, and the curve remains relatively normal. Inflationary or deflationary shocks shift yields across maturities, with the short end reacting first to changes in policy expectations.

As a result, different duration buckets respond to different economic forces:

- **1–3 years (short duration):** Primarily driven by near-term monetary policy, with low interest-rate sensitivity and relatively low volatility. Usually performs better in rising-rate environments.
- **3–5 years (short–intermediate):** Influenced by policy expectations and medium-term macro dynamics, offering a balanced risk–return profile.
- **5–10 years (intermediate):** Sensitive to inflation expectations, growth outlook, and term-premium movements. Performs well during slowdowns and easing cycles, but with higher volatility.
- **10+ years (long duration):** Most sensitive to yield changes, driven by long-run inflation expectations, real rates, and global risk sentiment. High potential returns in sustained easing regimes, but vulnerable during inflationary periods.

3 Factor Strategies on WGBI Index

We analyze duration-timing strategies on the WGBI using returns fully hedged into CHF. This removes exchange-rate effects that could dominate short-horizon returns and obscure the role of duration. It also reflects the perspective of a Swiss asset manager (like OLZ), for whom CHF-hedged returns are the relevant investment measure.

We evaluate factor-based strategies such as Carry, Value, and Momentum, together with simple combinations and transformations of these factors. At this stage, we deliberately exclude macroeconomic and broad market variables to isolate the predictive content of factor signals alone.

3.1 Carry, Value, and Momentum

a. Signal Definitions and Economic Motivation

In the following, we denote by $s_{b,t}^{f}$ the signal associated with factor f at time t for maturity bucket b. Let $y_{b,t}$ denote the yield of bucket b at time t, and let $r_{b,t}$ be the daily return of bucket b on day t. We write $\bar{y}_{b,t}^{(xd)}$ for the average yield over the past x days, and $r_{b,t}^{(xd)}$ for the cumulative return over the same x-day window. Similarly, throughout the superscripts d, w, m and y indicate day, week, month, and year horizons, respectively.

We consider the following standalone factor signals:

- **Carry**

  $$s_{b,t}^{\text{carry}} = \bar{y}_{b,t}^{(5d)}$$

  Higher yields imply higher expected carry and roll-down if the curve remains stable. This signal favours buckets with relatively higher yields and is equivalent to betting that the yield curve will not change over the next 21 days. Yields are smoothed using a 5-day moving average to reduce short-term noise and the influence of transient outliers.

- **Value**

  - *Rolling historical Value (Value)*

    $$s_{b,t}^{\text{value}} = y_{b,t} - \bar{y}_{b,t}^{(5y)}$$

    A yield above its five-year average suggests that the bucket is "cheap" relative to its recent history and its price may revert upward. This signal favours buckets whose yield level exceeds their own past average.

  - *NSS curve-based Value (NSS)*

    $$s_{b,t}^{\text{NSS}} = y_{b,t} - \hat{y}_{b,t}^{\text{NSS}}$$

    Here the observed yield is compared to the yield implied by a Nelson–Siegel–Svensson curve ($\hat{y}_{b,t}^{\text{NSS}}$) fitted daily to the full cross-section of maturities. The curve is estimated by least squares using the average life of each maturity

segment available (i.e. 1–3y, 3–5y, 5–7y, 7–10y, 10–15y, 15–20y, and 20+y) as a proxy for its time to maturity. A yield above the smooth NSS curve indicates that the bucket is "cheap" relative to the structural shape of the curve and may mean-revert through a price increase.

In essence, the two Value signals bet on mean reversion but capture different forms of mispricing. The rolling historical measure is a *time-series* signal: it assumes that each bucket has its own typical yield level and that deviations from this level will revert. The NSS-based measure is a *cross-sectional* signal: it assumes that the yield curve should be smooth across maturities and detects distortions relative to that shape.

- **Momentum signals**

  - *Directional Momentum (Mom1)*

    $$s_{b,t}^{\text{mom1}} = \text{sign}\left(r_{b,t}^{(1w)}\right) + \text{sign}\left(r_{b,t}^{(2w)}\right) + \text{sign}\left(r_{b,t}^{(1m)}\right)$$

    Captures the consistency of short-term returns but may become uninformative when all buckets move together. Indeed, in such regimes, all buckets may rise or fall together for extended periods, causing the directional signal to assign the same score to every bucket. This effectively collapses the strategy into an equal-weighted allocation, providing no useful cross-sectional information. For this reason, we introduce additional Momentum variants that extract trend strength over different horizons.

  - *1-month cumulative Momentum (Mom2)*

    $$s_{b,t}^{\text{mom2}} = r_{b,t}^{(1m)}$$

    A short-term trend signal that assigns magnitude rather than discrete votes.

  - *6-month cumulative Momentum (Mom3)*

    $$s_{b,t}^{\text{mom3}} = r_{b,t}^{(6m)}$$

    A medium-term trend indicator that reacts more slowly and captures smoother directional movements.

  - *Blended Momentum (Mom4)*

    $$s_{b,t}^{\text{mom4}} = \tfrac{1}{2}\, r_{b,t}^{(1m)} + \tfrac{1}{2}\, r_{b,t}^{(6m)}$$

    A combination of short- and medium-term horizons that stabilises the signal and reduces noise.

In all cases, the underlying idea is the same: we bet on a positive relationship between recent performance and future performance. These Momentum indicators are therefore purely time-series signals, designed to capture persistent trends within each duration bucket. Although they are constructed in a time-series manner, they are ultimately used to form cross-sectional allocations across buckets at each rebalancing date.

b. Portfolio Construction

To transform the signals into portfolio weights, we first standardise each factor cross-sectionally at every rebalancing date:

$$\tilde{s}^f_{b,t} = \frac{s^f_{b,t} - \mu^f_t}{\sigma^f_t}.$$

where μ^f_t and σ^f_t are the cross-sectional mean and standard deviation of the factor signal $s^f_{b,t}$ across buckets at rebalancing date t. This ensures that only the relative strength of each bucket's signal matters at time t, and provides a stable scale for mapping signals into portfolio weights.

Since all strategies are fully invested and long-only, and are rebalanced every 21 days as explained in subsection 1.1), we convert the signals into weights using a softmax transformation:

$$\omega^f_{b,t} = \frac{\exp\left(\tilde{s}^f_{b,t}/\tau\right)}{\sum_j \exp\left(\tilde{s}^f_{j,t}/\tau\right)},$$

where we set $\tau = 0.5$, a value that balances concentration and diversification in the resulting allocations and performs well on the training sample. At this stage, τ is kept fixed to focus on the baseline behaviour of the factor signals. In the multi-country extension, where a richer cross-section is available, we allow τ to vary and analyse its effect on portfolio concentration and performance.

We adopt the softmax transformation rather than simple rescaling for three reasons. First, softmax produces strictly positive weights, which is appropriate given the long-only nature of the strategies. Alternative specifications, such as normalising by the sum of signals, would require clipping negative signals to zero, thereby discarding potentially informative cross-sectional variation and, in extreme cases, concentrating the entire allocation in a single bucket when all other signals are negative. Second, softmax preserves relative signal magnitudes: the distance between signals matters, so a bucket that is only weakly favoured does not mechanically receive an extreme allocation. Third, more sophisticated allocation rules cannot be applied at this stage, since the signals are not direct return forecasts. In particular, mean–variance or related optimisation frameworks are not applicable here, as they require explicit expectations of returns and covariances rather than relative ranking signals.

Finally, the purpose of this section is to isolate the contribution of signals derived from factor exposures and PCA-related information, rather than to optimise the signal-to-portfolio mapping itself. The portfolio construction step is therefore kept deliberately simple and transparent, so that differences in performance can be attributed to the informational content of the signals rather than to the choice of allocation rule.

c. Results

We assess the performance of these signals using Table 2. We also note that the results obtained in the evaluation periods and presented here (or in the following paragraph) are not used in any way when combining strategies in the next section. We model the cost of turnover as explained at the end of subsection 1.1 with $c = 5$ bps.

Training Sample (2005–2013)

Strategy	Cum.	Sharpe	Sharpe/turn
WGBI 1–3y	0.134	1.807	1.807
WGBI 3–5y	0.228	1.331	1.331
WGBI 5–10y	0.338	1.071	1.071
WGBI 10+y	0.405	0.674	0.674
Equal-weighted	0.275	1.009	1.008
Carry	0.388	0.696	0.694
Value	0.320	1.102	1.082
NSS	0.169	0.873	0.789
Momentum	0.255	0.984	0.780
Momentum2	0.388	0.947	0.779
Momentum3	0.298	0.731	0.652
Momentum4	0.346	0.845	0.724

Evaluation Period I (2014–2019)

Strategy	Cum.	Sharpe	Sharpe/turn
WGBI 1–3y	−0.035	−1.193	−1.193
WGBI 3–5y	0.020	0.249	0.249
WGBI 5–10y	0.137	0.835	0.835
WGBI 10+y	0.403	0.989	0.989
Equal-weighted	0.121	0.784	0.783
Carry	0.362	0.993	0.986
Value	0.038	0.549	0.500
NSS	0.129	0.771	0.678
Momentum	0.196	0.979	0.838
Momentum2	0.362	1.276	1.140
Momentum3	0.247	0.787	0.738
Momentum4	0.279	0.919	0.833

Evaluation Period II (2019–2024)

Strategy	Cum.	Sharpe	Sharpe/turn
WGBI 1–3y	−0.049	−0.666	−0.666
WGBI 3–5y	−0.070	−0.429	−0.429
WGBI 5–10y	−0.084	−0.300	−0.300
WGBI 10+y	−0.208	−0.389	−0.389
Equal-weighted	−0.102	−0.410	−0.410
Carry	−0.077	−0.202	−0.217
Value	−0.098	−0.590	−0.624
NSS	−0.083	−0.284	−0.306
Momentum	−0.111	−0.415	−0.510
Momentum2	−0.108	−0.340	−0.476
Momentum3	−0.041	−0.128	−0.193
Momentum4	−0.094	−0.288	−0.387

Full Evaluation Sample (2014–2024)

Strategy	Cum.	Sharpe	Sharpe/turn
WGBI 1–3y	−0.081	−0.792	−0.792
WGBI 3–5y	−0.063	−0.259	−0.259
WGBI 5–10y	0.004	0.027	0.027
WGBI 10+y	0.015	0.055	0.055
Equal-weighted	−0.027	−0.055	−0.056
Carry	0.198	0.318	0.307
Value	−0.077	−0.319	−0.357
NSS	−0.011	−0.007	−0.060
Momentum	0.008	0.037	−0.069
Momentum2	0.059	0.128	0.001
Momentum3	0.083	0.172	0.110
Momentum4	0.057	0.125	0.030

Table 2: Performance of pure factor strategies across training and out-of-sample periods. The table reports cumulative returns, Sharpe ratios, and turnover-adjusted Sharpe ratios for WGBI maturity buckets and characteristic-based strategies over the training sample (2005–2013), two evaluation subperiods (2014–2019 and 2019–2024), and the full out-of-sample period (2014–2024).

Across the training period (2005–2013), all strategies appear to perform well, but this outcome must be interpreted with care. The results are structurally biased because the signals were designed and selected using this very window, and also because the regime itself is exceptionally favourable: yields decline almost monotonically for a decade, lifting returns across the entire curve. In such an environment even a naïve equal-weighted portfolio achieves a relatively good Sharpe ratio, which already indicates that limited timing skill is required to perform well.

In the 2014–2019 window, performance becomes more heterogeneous. Long-duration buckets lead the market, and Momentum (especially Momentum2) continues to perform well, whereas Value and NSS weaken materially. Their mean-reversion logic fails in a regime where yields move smoothly in one direction. Carry also remains effective due

to its systematic tilt toward higher-yielding segments.

The 2019–2024 period is characterized by inflation shocks and rapid tightening cycles. All duration buckets generate negative returns, and most factor signals follow the same pattern. Momentum breaks down as trends repeatedly reverse; Value and NSS remain weak throughout. Only long-horizon Momentum (Momentum3) shows limited relative resilience, though still negative overall. In a broad rates sell-off, long-only duration strategies struggle to produce positive returns, so a meaningful outcome can simply be to design a strategy that reduces the drawdown.

Over the full evaluation horizon (2014–2024), combining the benign early subperiod with the highly adverse later years, most signals average out to flat or mildly negative performance. Carry is the best performing factor strategy (in terms of Sharpe ratio and cumulative returns), while long-horizon Momentum strategy again appears the least fragile (with the highest Sharpe ratio over the 2019–2024 period). The returns and allocations of these two strategies are shown in detail in Figures 1 and 2.

Across the training window (2005–2013), the strong performance of Value and Momentum should be viewed in the context of the regime itself. Much of the empirical literature documenting the strength of these signals was built on long samples dominated by yield declines. The training period here shares exactly that structure, so the strong performance of Value and Momentum is unsurprising and partly reflects the same historical bias: these signals were never stress-tested against a decade like the post-2014 period.

That said, even after accounting for regime effects, several conclusions remain robust. Carry and Momentum strategies outperform the equal-weighted benchmark in the training window (in terms of cumulative returns) and continue to dominate in the more challenging evaluation period, both in terms of returns and Sharpe ratio. These strategies allocate capital more effectively across the yield curve, mitigate the limitations of static schemes, and preserve relative performance even when the overall duration environment becomes unfavourable. As a result, Carry and Momentum emerge as the most reliable pure-factor approaches and are clearly preferable to a simple equal-weighted allocation.

Portfolio turnover is generally low across strategies because the underlying signals evolve smoothly over time and do not change abruptly, which keeps rebalancing activity and transaction costs contained. Turnover costs therefore have a present but moderate impact and do not materially affect the overall conclusions, also given the limited level of assumed transaction costs. Momentum-based strategies are more affected by turnover because the signal evolves more rapidly over time, leading to more frequent reallocations, as can be seen from the second row of Figure 2, and resulting in a deterioration of their net Sharpe ratio.

Figure 1: Top row: cumulative returns of the Carry portfolio and the four WGBI buckets plus the equal–weighted benchmark. Bottom row: portfolio weights, showing persistent allocation to the highest–yielding bucket except when relative yields compress or invert.

Figure 2: Top row: cumulative returns of Momentum3 portfolio versus benchmarks. Bottom row: portfolio weights implied by the 6-month Momentum signal, with slow regime shifts and delayed reactions around turning points.

3.2 Combination of Signals

As discussed in subsection 3.1, each factor reflects a distinct economic mechanism. Carry assumes the yield curve remains locally stable; Value exploits mean reversion in yield levels; Momentum captures the empirical persistence of recent price trends.

For the composite model, we retain the following three signals:

- *Carry*;
- *Value* (rolling historical version, not NSS, since it delivers stronger performance in the training sample and is more robust);
- *Momentum2* (1-month cumulative return), which we select in place of the directional-vote Momentum or the long-horizon variants. Momentum1 suffers from the issue that it might assign identical signals to all maturities and Momentum2 offers the most attractive combination of Sharpe ratio and cumulative return in the training window.

a. Equal-weighted Composite

The baseline combination simply averages the three standardised signals:

$$s_{b,t}^{\text{total}} = \frac{1}{3}\Big(\tilde{s}_{b,t}^{\text{carry}} + \tilde{s}_{b,t}^{\text{value}} + \tilde{s}_{b,t}^{\text{mom2}}\Big).$$

This procedure allows the model to automatically favour whichever signal is strongest at a given time.

b. Adaptive Combination via Information Coefficients

The equal-weighted composite assumes that the economic interpretation of each factor is stable through time. However, this need not be the case. For example, Value relies on mean reversion, but during persistent trending regimes deviations from the average may actually widen rather than revert. To account for these possible sign reversals or regime dependencies, we build an adaptive weighting scheme using daily *Information Coefficients* (which we will refer to as ICs). The procedure consists of the following steps:

1. **Forward 21-day returns.** For each bucket b, compute the forward cumulative 21-day return:
$$r_{b,t}^{(21d)} = \prod_{k=1}^{21}(1 + r_{b,t+k}) - 1.$$
2. **Daily cross-sectional IC.** The predictive ability of each factor f is measured by the daily cross-sectional (Pearson) correlation between its signal and subsequent returns:
$$IC_t^f = \text{corr}\big(s_{b,t}^f,\, r_{b,t}^{(21d)}\big).$$
3. **Rolling IC averages.** To extract persistent information and reduce noise, compute rolling averages (e.g., over 100 days):
$$\overline{IC}_t^f = \frac{1}{N}\sum_{i=1}^{N} IC_{t-i}^f.$$

4. **IC-based factor weights.** Normalise the averaged ICs so that the composite depends on each factor's relative predictive strength:

$$\varepsilon_t^f = \frac{\overline{IC}^f_{t-21}}{\sum_g |\overline{IC}^g_{t-21}|}.$$

Note that weights used at time t are computed from ICs available up to $t - 21$, to avoid clairvoyance.

5. **Adaptive composite signal.** For each bucket, form an IC-weighted composite:

$$s^{IC}_{b,t} = \varepsilon_t^{\text{carry}}\, \tilde{s}^{\text{carry}}_{b,t} + \varepsilon_t^{\text{value}}\, \tilde{s}^{\text{value}}_{b,t} + \varepsilon_t^{\text{mom2}}\, \tilde{s}^{\text{mom2}}_{b,t}.$$

c. Combination via Logistic Regression

Some predictive content may lie in the short-term dynamics of Value, Momentum, and Carry themselves. Their recent changes (for example over 5 or 20 days) can contain information that is not visible in the raw levels, and as noted earlier certain signals may relate inversely to subsequent returns. For this reason we also test a Logistic Regression with light feature engineering. The features used are deliberately simple and constructed only from the two extreme buckets (1–3y and 10y+): the raw relative spreads $\text{Carry}_{10y+} - \text{Carry}_{1\text{–}3y}$, $\text{Value}_{10y+} - \text{Value}_{1\text{–}3y}$, $\text{Mom}_{10y+} - \text{Mom}_{1\text{–}3y}$, together with their short-term changes computed over 5 days and over 20 days. These quantities capture both the cross-sectional gap between the long- and short-duration buckets and the recent evolution of that gap.

The model is trained to predict which of the two duration buckets will outperform over the subsequent 21 days. The portfolio weights are then obtained directly from the model's output probabilities, so that the allocation reflects the estimated likelihood that each bucket will deliver the higher forward return.

A broader exploration of feature sets and more sophisticated machine learning models is intentionally left for later, since the aim of this section is to evaluate the baseline predictive content of the underlying factors. We apply Logistic Regression only to the two extreme buckets, the lowest-duration and the highest-duration portfolios, rather than using a multinomial model across all four buckets. The rationale for this choice is discussed in Section 4.2 as the details of the training.

d. Results

We now evaluate the performance of these strategies by analyzing Table 3. We also include Carry, Value, and Momentum2 as reference points for combined strategies.

Training Sample (2005–2013)

Strategy	Cum.	Sharpe	Sharpe/turn
Equal-weighted	0.275	1.009	1.008
Carry	0.388	0.696	0.694
Value	0.320	1.102	1.082
Momentum2	0.388	0.947	0.779
Signal Comb	0.421	1.028	0.866
IC	0.320	0.825	0.743
Logistic Reg	0.210	0.574	0.499

Evaluation Period I (2014–2019)

Strategy	Cum.	Sharpe	Sharpe/turn
Equal-weighted	0.121	0.784	0.783
Carry	0.362	0.993	0.986
Value	0.038	0.549	0.500
Momentum2	0.362	1.276	1.140
Signal Comb	0.276	1.448	1.261
IC	0.246	0.859	0.781
Logistic Reg	0.227	0.887	0.802

Evaluation Period II (2019–2024)

Strategy	Cum.	Sharpe	Sharpe/turn
Equal-weighted	−0.102	−0.410	−0.410
Carry	−0.077	−0.202	−0.217
Value	−0.098	−0.590	−0.624
Momentum2	−0.108	−0.340	−0.476
Signal Comb	−0.085	−0.348	−0.456
IC	0.007	0.048	−0.047
Logistic Reg	−0.117	−0.344	−0.389

Full Evaluation Sample (2014–2024)

Strategy	Cum.	Sharpe	Sharpe/turn
Equal-weighted	−0.027	−0.055	−0.056
Carry	0.198	0.318	0.307
Value	−0.077	−0.319	−0.357
Momentum2	0.059	0.128	0.001
Signal Comb	0.095	0.222	0.098
IC	0.115	0.256	0.164
Logistic Reg	−0.020	−0.014	−0.075

Table 3: Performance of combined factor strategies across training and out-of-sample periods. The table reports cumulative returns, Sharpe ratios, and turnover-adjusted Sharpe ratios for the equal-weighted benchmark and selected combined-signal approaches over the training sample (2005–2013), two evaluation subperiods (2014–2019 and 2019–2024), and the full out-of-sample period (2014–2024).

Signal Combined shows a stable and intuitive pattern across all periods. In the training sample it performs close to the strongest individual signals, with Sharpe just below Momentum2 but above Carry and Value. In 2014–2019, it improves further: the net Sharpe is 1.26, broadly in line with Momentum2 and much higher than Value and Carry. In the 2019-2024 period, it is able to limit the losses compared to some of the single factors. This confirms that averaging the three signals produces a robust profile without depending on a single factor.

IC behaves differently because its weights are recalibrated based on which factor is actually predictive in each window. In the training and early-evaluation periods it is slightly weaker than Signal Combined but still competitive, with net Sharpe ratios around 0.74 (training) and 0.78 (2014–2019). The important point is the 2019–2024 cycle: this is the only sample where every duration bucket and every static factor turns negative. Here Signal Combined also suffers (net Sharpe –0.456), while IC remains slightly positive in terms of gross Sharpe (0.048), although turnover costs reduce the net Sharpe to –0.047. The table makes this clear: IC is the only approach with a positive gross Sharpe in the sell-off regime.

Over the full 2014–2024 sample, the two approaches converge in terms of overall performance. Signal Combined attains a gross Sharpe of 0.222 and a net Sharpe of 0.098, while the IC-based strategy improves to 0.256 gross and 0.164 net, outperform-

Figure 3: IC–based composite strategy on WGBI duration buckets. Top row: cumulative returns versus benchmarks. Second row: portfolio weights implied by the adaptive IC signal. Third row: violin plot of IC–weighted signals across buckets. Bottom row: time series of rolling information coefficients for Carry, Value, and Momentum, which determine the dynamic factor weights used in the composite signal.

ing Value and Momentum2 but remaining below Carry. This indicates that (i) static diversification across Carry, Value, and Momentum already provides a strong and stable baseline, and (ii) the IC-based mechanism adds resilience in stressed periods by progressively down-weighting underperforming signals and reallocating toward those that retain predictive content. Because factor weights are computed from rolling averages of ICs, the IC approach features smoother adjustments over time, which helps contain turnover with respect to Signal Combined.

We can use the third and fourth row of Figure 3 to diagnose the predictive performance of Value, Carry and Momentum. Carry exhibits consistently positive information coefficients across all periods, confirming that its relative-yield signal maintains a sta-

ble relation with forward returns. Momentum also remains mostly positively correlated with future performance, with only moderate variation in stressed regimes. Value, by contrast, flips sign repeatedly: in some periods it correlates positively, in others negatively. This instability indicates that Value is not a reliable standalone predictor and that its forecasting direction is regime-dependent.

The time-series of rolling ICs in the fourth row reinforces the same point. Carry's IC stays above zero almost everywhere, Momentum fluctuates more but remains predominantly positive, especially during trending phases. Value, however, oscillates around zero with extended negative stretches, reflecting reversals in its economic interpretation. These dynamics justify the need for an adaptive mechanism like the very IC-based weighting.

Finally, the Logistic Regression behaves worse than one would expect. By construction, one would anticipate that a model that updates weights based on recent factor behaviour should at least limit drawdowns relative to static combinations, yet this is not what we observe. The strategy fails to protect capital in the 2019–2024 regime and does not deliver a clear improvement in the earlier periods either. This suggests that, in its current form, the machine learning layer is either too weak to extract additional structure from the inputs, or it is overfitting noise given how light the feature set is. In the next section we will see that the main issue is not the use of Logistic Regression per se, but the lack of sufficiently rich and informative features.

3.3 Predictive Power of PCA

In this subsection, we test whether principal components extracted from the yield curve contain incremental predictive information. As established in the literature (e.g., [9]), PCA applied to yields typically produces factors interpreted as *level*, *slope*, and *curvature*. These components summarise the geometry of the term structure and provide a compact description of whether the curve is normal, inverted, or humped.

To exploit this information, we compute principal components using a rolling window of one year (approximately 250 trading days) of yield data. At each rebalancing date t, the yield vector observed at $t-1$ (on the previous day) is projected onto the principal components estimated over the corresponding rolling window, yielding time-varying scores for the level, slope, and curvature factors. We deliberately rely on a short rolling window rather than the full available history, as the objective is to capture short-horizon changes in the shape of the yield curve rather than long-run average yield levels.

Interpretation of the PCA scores is standard:

- a negative PC2 is typically associated with an inverted curve (short rates above long rates);
- a positive PC3 is indicative of a humped curve, where intermediate maturities yield more than both short and long ends.

An alternative would have been to fit the NSS curve at each date using the available information, but we did not explore this option due to practical and computational

constraints, and because it provides only point-in-time information without linking the curve's shape to past observations, making it not directly comparable to our approach.

a. Modification of Carry

We use the information extracted from PCA to construct a simple modification of the Carry signal. Whenever the second principal component falls below -0.2, a threshold selected on the training sample, we invert the sign of the Carry signal. This threshold is chosen to identify episodes of pronounced curve inversion rather than mild flattening: PC2 values below this level correspond to regimes in which the standard Carry logic is known to break down. The results are not sensitive to the exact choice of the threshold; similar performance is obtained over a range of similar negative values, indicating that the strategy relies on detecting a clear inversion regime rather than on fine-tuning the cutoff.

The economic intuition is that a normally sloped yield curve tends to persist, whereas inversions typically emerge during stress episodes and represent an unstable configuration that policymakers often attempt to reverse. By inverting Carry in such regimes, the strategy explicitly positions for a re-normalisation of the yield-curve slope.

b. PCA Threshold Strategy

This second PCA–based strategy uses the shape of the yield curve to guide duration allocation, without relying on traditional factor signals. At each rebalancing date, we observe the values of PC2 and PC3 (computed from the previous one year of yield data) and classify the yield curve into three simple regimes:

- **Humped curve (**$\mathrm{PC3} > 0.2$**).** When the curve is humped, intermediate maturities typically offer the most attractive risk–return trade-off. The strategy allocates 50% to the 3–5y bucket and 50% to the 5–10y bucket.
- **Inverted curve** ($\mathrm{PC2} < -0.2$ and $\mathrm{PC3} < 0.2$). Under inversion, short maturities dominate: they carry higher yields and usually benefit most when rates eventually decline. The portfolio is fully allocated to the 1–3y bucket.
- **Normal curve (otherwise).** When the curve is upward sloping, long maturities generally deliver the strongest Carry. The portfolio allocates entirely to the 10y+ bucket.

In essence, this strategy converts PCA information into a *curve–shape switch*: depending on whether the yield curve is normal, inverted, or humped, duration exposure is shifted toward the segment that historically performs best in that environment. This results in a simple and highly interpretable regime–based allocation rule driven directly by the geometry of the term structure.

We then extend this PCA–regime strategy by allowing the portfolio weights within each curve configuration to be parametrised rather than fixed. Each regime is associated with a vector of weights of the form:

- *Humped curve* (PC3 > 0.2): weights $[0,\ \alpha,\ 1-\alpha,\ 0]$.
- *Inverted curve* (PC2 < -0.2 and PC3 < 0.2): weights $[\gamma,\ 1-\gamma,\ 0,\ 0]$.
- *Normal curve* (otherwise): weights $[0,\ 0,\ 1-\delta,\ \delta]$.

A grid search over (α, γ, δ) is performed on the training sample. For each parameter triplet, the full strategy is simulated and its Sharpe ratio (optionally adjusted for turnover) is computed. The parameter set that maximises performance is retained. In the tables below, we refer to the unoptimised and optimised versions of this approach as *Full PCA* and *Full PCA**, respectively.

c. Logistic Regression with Yield Curve Information

We expect signals to exhibit similar behaviour when the yield curve is in a comparable state. To exploit this, we split the sample into two regimes based on the sign and magnitude of the second principal component. PCA is computed exactly as before, and at each rebalancing date we classify the current curve using the previous day's PCA values. We then train the Logistic Regression on a restricted subsample consisting only of days in which the curve was in the same regime. Concretely, if today's state satisfies PC2 < -0.2, the model is trained only on past observations with PC2 < -0.2; the same logic applies symmetrically when the curve is on the opposite side.

d. Results

We now evaluate the performance of these strategies using Table 4. We also include Carry and Logistic Regression strategies as reference points for comparison.

The Modified Carry strategy (Figure 4) is consistently more robust than the standard Carry approach. Within the training window it delivers a higher Sharpe ratio; in the first evaluation period it remains competitive or superior; and, most importantly, during the 2019–2024 sell-off regime it substantially mitigates losses relative to plain Carry. The PCA-based adjustment to curve shape enables the strategy to avoid the characteristic breakdown of Carry when the yield curve flattens or inverts, resulting in markedly more stable performance across regimes. This modification entails higher turnover, as duration exposure can shift abruptly when the threshold is crossed. Nevertheless, even after accounting for transaction costs, the strategy is on par with the standard Carry in terms of net Sharpe over the full evaluation window.

PCA-based strategies deliver the high full-sample gross Sharpe ratio over 2014–2024, but costs reduce their net performance. However, their weakness is evident in the second evaluation window: both specifications react too slowly to the rapid rate-hiking cycle and experience deep drawdowns. This shows that, although highly informative on average, they are not robust in fast macro rotations.

Despite this, the underlying PCA signal remains powerful: it cleanly captures level, slope, and curvature dynamics and provides structural predictive information about the term structure. The poor short-horizon robustness does not diminish its value; rather, it suggests that a PCA-driven signal can serve as a foundation for more sophisticated and potentially superior strategies. Such strategies, however, tend to incur high turnover

Training Sample (2005–2013)

Strategy	Cum.	Sharpe	Sharpe/turn
Equal-weighted	0.275	1.009	1.008
Carry	0.388	0.696	0.694
Carry mod	0.428	0.929	0.848
Full PCA	0.383	0.898	0.759
Full PCA*	0.378	0.959	0.824
Logistic Reg	0.210	0.574	0.499
Log Reg PCA	0.238	0.623	0.536

Evaluation Period I (2014–2019)

Strategy	Cum.	Sharpe	Sharpe/turn
Equal-weighted	0.121	0.784	0.783
Carry	0.362	0.993	0.986
Carry mod	0.359	1.226	1.096
Full PCA	0.298	1.105	0.932
Full PCA*	0.268	1.109	0.939
Logistic Reg	0.227	0.887	0.802
Log Reg PCA	0.294	1.116	1.021

Evaluation Period II (2019–2024)

Strategy	Cum.	Sharpe	Sharpe/turn
Equal-weighted	−0.102	−0.410	−0.410
Carry	−0.077	−0.202	−0.217
Carry mod	−0.025	−0.049	−0.105
Full PCA	−0.144	−0.329	−0.399
Full PCA*	−0.132	−0.333	−0.400
Logistic Reg	−0.117	−0.344	−0.389
Log Reg PCA	−0.197	−0.604	−0.666

Full Evaluation Sample (2014–2024)

Strategy	Cum.	Sharpe	Sharpe/turn
Equal-weighted	−0.027	−0.055	−0.056
Carry	0.198	0.318	0.307
Carry mod	0.220	0.383	0.294
Full PCA	0.238	0.394	0.279
Full PCA*	0.191	0.352	0.241
Logistic Reg	−0.020	−0.014	−0.075
Log Reg PCA	0.038	0.093	0.020

Table 4: Performance of PCA-based strategies across training and out-of-sample periods. The table reports cumulative returns, Sharpe ratios, and turnover-adjusted Sharpe ratios for the equal-weighted benchmark, Carry-based baselines, and PCA-enhanced variants over the training sample (2005–2013), two evaluation subperiods (2014–2019 and 2019–2024), and the full out-of-sample period (2014–2024).

costs, as they can switch abruptly across duration buckets. A natural extension of this work would be to study their performance under weight-smoothing schemes, for instance by introducing gradual rebalancing.

The optimised PCA strategy performs worse than the naïve PCA version because, ex post, we know that regimes change abruptly. This reinforces the point that robustness matters more than squeezing out marginal in-sample gains. A strategy tuned too closely to historical conditions collapses as soon as the environment shifts, whereas a more restrained, less overfit specification degrades more slowly and remains usable across regimes.

PCA-based strategies perform less well in the U.S. Treasury market when used on their own. Allocations driven solely by PCA factors tend to underperform the benchmark across most periods, even before transaction costs, indicating that latent yield-curve components have limited standalone timing power. By contrast, logistic regression performs better, particularly when PCA factors are used as inputs and after accounting for transaction costs. This suggests that PCA is more effective as an auxiliary feature in a supervised framework, rather than as the basis of a standalone timing strategy.

Finally, Logistic Regression with PCA performs slightly better than the standard logistic model in the training sample and over the full 2014–2024 period. The improve-

Figure 4: Top row: cumulative returns of the Carry modified portfolio and the four WGBI buckets plus the equal–weighted benchmark. Bottom row: portfolio weights of the Carry-modified strategy. Unlike standard Carry, the allocation shifts frequently and tracks the regime indicated by PC2.

ment is small, and the Sharpe ratios should be read cautiously: when returns hover around zero, minor changes in volatility can move the Sharpe up or down without indicating real predictive power. The method is not robust in 2019–2024, where performance drops sharply. The post-pandemic sell-off regime has few precedents in the historical sample, so a regime-based classifier has little reliable structure to learn from.

Performance deteriorates markedly in the 2019–2024 period, where the model fails to adapt to the rapid rate-hiking cycle. The post-pandemic sell-off represents a regime with few historical precedents in the sample, leaving a regime-based classifier with limited informative structure to exploit.

3.4 Sharpe Ratio Tests for Factor Strategies

In Table 5 we report HAC-based and bootstrap p-values for Sharpe ratio differentials relative to the equal-weighted benchmark in the WGBI universe. As discussed in detail in Subsection 1.1, inference is conducted using heteroskedasticity- and autocorrelation-consistent standard errors (HAC) together with a bootstrap procedure (Boot), explicitly accounting for serial dependence and non-normality in excess returns. In the same section, we also explain why we do not apply multiple-testing corrections.

The results indicate that only Carry and Modified Carry achieve clear statistical significance. Under both inference procedures, their p-values lie below the 5% threshold, providing consistent evidence that their risk-adjusted outperformance is unlikely to be driven by sampling variation. This supports the interpretation that carry exploits genuine cross-country yield-curve dispersion.

Full PCA and its constrained variant lie at the boundary of conventional significance levels. Their p-values around 5% suggest marginal evidence of outperformance,

Strategy	p_{HAC}	p_{Boot}
Carry	0.025	0.033
Carry Mod	0.039	0.045
Full PCA*	0.053	0.042
Full PCA	0.053	0.054
IC	0.146	0.143
Momentum3	0.207	0.216
Signal Combined	0.247	0.273
Log Reg Regime	0.330	0.354
Momentum4	0.333	0.321
Momentum2	0.389	0.418
NSS	0.534	0.526
Momentum	0.541	0.559
Log Reg	0.547	0.531
Value	0.927	0.925

Table 5: HAC-based and bootstrap p-values for the WGBI full evaluation sample (2014-2024), ordered by ascending HAC p-values. The null hypothesis is that each strategy does not outperform the equal-weighted benchmark in risk-adjusted terms (after transaction costs).

indicating that global term-structure information contains exploitable structure, but in a weaker and more model-dependent form than carry.

For the IC strategy, the outcome is more unexpected. Given its relatively high Sharpe ratio and its strong theoretical foundation one would have anticipated stronger statistical confirmation. Instead, both HAC and bootstrap p-values remain slightly above the conventional rejection threshold at 10%. This implies that, despite solid economic intuition and favourable full-sample performance, the variability of excess returns prevents formal statistical rejection of the null. The strategy appears economically meaningful but statistically less strong. Nevertheless, the results inthe table clearly indicate that IC dominates the Signal Combined approach. Its p-values are consistently lower under both inference procedures, reinforcing the view that weighting signals according to their correlation with subsequent returns is more effective than equally aggregating individual signals into a composite score.

All remaining strategies, including Value, NSS, the Momentum variants, the combined signal, and Logistic Regression specifications, display high p-values. Their Sharpe improvements are not statistically distinguishable from the benchmark.

Overall, the inference isolates carry as the only robust and statistically defensible factor in the global setting, while the majority of alternative signals, including theoretically well-motivated ones such as IC, fail to deliver statistically significant Sharpe ratio improvements once proper inference is applied.

3.5 Comments on Factor Strategies

Factor strategies show clear predictive content: several of them outperform the benchmark in economic terms, even if most do not generate statistically strong Sharpe ratios once formally tested. As shown by the HAC and bootstrap Sharpe ratio tests, only a

subset of strategies achieves conventional significance levels, while others display economically meaningful but statistically fragile improvements. With only yields and past returns as inputs, achieving a positive and robust Sharpe during broad fixed-income sell-offs is extremely difficult; both the performance tables and the inference results indicate that only the IC approach consistently maintains favourable risk-adjusted behaviour in stress periods. Momentum appears explicitly as a signal and implicitly inside both logistic models and the IC procedure, where recent observations directly shape forecasts. Although the standalone momentum variants are not statistically significant in the Sharpe tests, their presence within the better-performing combined frameworks suggests that they contribute economically relevant information.

The last two sections above highlight that PCA-based strategies carry meaningful predictive power. PC2, in particular, captures the shape of the curve and underlying regime shifts. The Sharpe ratio tests show that pure PCA allocations deliver p-values close to the 5% threshold, which in this setting can be regarded as economically meaningful and statistically strong evidence, given the limited sample size and the dependence structure of returns.

From a practical allocation perspective, and in light of the formal Sharpe ratio tests, the evidence points toward selecting either the Carry-based specifications, given their statistically significant outperformance, or the IC approach, which, despite falling marginally short of conventional thresholds, combines a solid theoretical foundation with comparatively robust empirical performance.

4 Strategies on the U.S. Government Bond Index

In this section, we examine duration-timing strategies in the U.S. Treasury market. Returns are expressed in U.S. dollars in order to abstract from currency effects. We first assess the performance of the factor signals introduced in the previous section. Relative to the global WGBI universe, where the U.S. represents approximately 44% of the index, the U.S. Treasury market is more informationally efficient, with yields reflecting a richer set of macroeconomic information. As a result, purely yield-curve-based factor signals are expected to exhibit weaker and less stable performance in this setting. This motivates extending the analysis by incorporating macroeconomic and market variables into the duration-timing framework. Transaction costs are set to 5 bps per unit of turnover, consistent with the assumptions used in the previous sections. The benchmark is an equal-weighted portfolio across the four U.S. Treasury maturity buckets.

4.1 Analysis of Factor Strategies

By inspecting Table 6, the shortest-duration bucket emerges as the best-performing segment in the U.S. data, displaying the highest Sharpe ratio. At first sight this contrasts with the WGBI results, where the shortest bucket seemed to be the weakest. The comparison is misleading: in both datasets the shortest bucket follows a largely monotonic path, but in the WGBI case the appreciation of the Swiss franc suppresses the observed returns on low-duration bonds (since we were using returns fully hedged in CHF), and their extremely low volatility makes the Sharpe ratio appear artificially unfavourable. In reality, the shortest bucket is exactly the segment that preserves capital during broad fixed-income sell-offs in both setups.

Conversely, in the Evaluation Sample the strong performance of the shortest bucket can obscure the true contribution of timing strategies: a naïve rule that always selects the lowest-duration bucket would look strong on the full sample while delivering little value in terms of active allocation. Since the objective is to maximise returns for a given level of risk, one cannot evaluate a timing strategy by inspecting the Sharpe ratio in isolation; one must control for the structural dominance of the low-duration bucket before assessing whether a strategy genuinely extracts information rather than merely reverting to the safest segment.

Given this, the only single-factor strategies that appear to outperform the benchmark in Sharpe terms are Value and NSS. This outperformance, however, is largely mechanical: both strategies generate an almost static allocation concentrated in the shortest-duration buckets, as illustrated for NSS in Figure 5, with Value exhibiting a very similar pattern. Such positioning provides protection during sell-offs but prevents the strategies from capturing returns in Evaluation Period I, when longer durations dominate performance. As a result, these approaches do not represent genuinely investable signals; they behave instead like static underweights to duration rather than adaptive timing strategies.

Training Sample (2005–2013)

Strategy	Cum.	Sharpe	Sharpe/turn
U.S. 1–3y	0.273	1.888	1.888
U.S. 3–5y	0.443	1.147	1.147
U.S. 5–10y	0.568	0.830	0.830
U.S. 10+y	0.650	0.512	0.512
Equal-weighted	0.487	0.814	0.813
Carry	0.635	0.547	0.543
Carry Mod	0.342	0.468	0.398
Value	0.583	0.713	0.706
NSS	0.691	0.675	0.638
Momentum	0.488	0.799	0.709
Momentum2	0.600	0.723	0.636
Momentum3	0.404	0.460	0.434
Momentum4	0.565	0.606	0.571
Signal Comb	0.626	0.724	0.661
IC	0.541	0.650	0.613
Log Reg	0.336	0.459	0.420
Log Reg PCA	0.401	0.537	0.491
Full PCA	0.258	0.342	0.267
Full PCA*	0.300	0.422	0.342

Evaluation Period I (2014–2019)

Strategy	Cum.	Sharpe	Sharpe/turn
U.S. 1–3y	0.077	1.422	1.422
U.S. 3–5y	0.133	0.878	0.878
U.S. 5–10y	0.217	0.775	0.775
U.S. 10+y	0.542	0.683	0.683
Equal-weighted	0.236	0.791	0.790
Carry	0.488	0.686	0.685
Carry Mod	0.403	0.851	0.805
Value	0.181	1.135	1.112
NSS	0.195	1.087	0.992
Momentum	0.229	0.659	0.567
Momentum2	0.391	0.755	0.672
Momentum3	0.149	0.297	0.275
Momentum4	0.152	0.306	0.271
Signal Comb	0.255	0.722	0.632
IC	0.310	0.624	0.589
Log Reg	0.364	0.735	0.692
Log Reg PCA	0.390	0.816	0.765
Full PCA	0.255	0.519	0.440
Full PCA*	0.211	0.502	0.419

Evaluation Period II (2019–2024)

Strategy	Cum.	Sharpe	Sharpe/turn
U.S. 1–3y	0.110	1.023	1.023
U.S. 3–5y	0.078	0.347	0.347
U.S. 5–10y	0.039	0.131	0.131
U.S. 10+y	−0.121	−0.055	−0.055
Equal-weighted	0.030	0.106	0.106
Carry	0.019	0.085	0.080
Carry Mod	0.038	0.110	0.071
Value	0.055	0.212	0.191
NSS	0.076	0.344	0.280
Momentum	−0.006	0.021	−0.057
Momentum2	−0.083	−0.090	−0.161
Momentum3	0.123	0.241	0.224
Momentum4	0.079	0.176	0.150
Signal Comb	0.092	0.210	0.157
IC	0.263	0.538	0.501
Log Reg	0.132	0.252	0.230
Log Reg PCA	0.099	0.215	0.178
Full PCA	−0.101	−0.071	−0.112
Full PCA*	−0.071	−0.051	−0.095

Full Evaluation Sample (2014–2024)

Strategy	Cum.	Sharpe	Sharpe/turn
U.S. 1–3y	0.154	0.960	0.960
U.S. 3–5y	0.161	0.437	0.437
U.S. 5–10y	0.177	0.294	0.294
U.S. 10+y	0.179	0.175	0.175
Equal-weighted	0.182	0.291	0.290
Carry	0.360	0.301	0.296
Carry Mod	0.264	0.305	0.260
Value	0.159	0.376	0.355
NSS	0.231	0.573	0.496
Momentum	0.212	0.304	0.227
Momentum2	0.218	0.233	0.165
Momentum3	0.141	0.173	0.153
Momentum4	0.186	0.212	0.184
Signal Comb	0.128	0.205	0.123
IC	0.312	0.357	0.323
Log Reg	0.297	0.310	0.278
Log Reg PCA	0.465	0.454	0.410
Full PCA	0.271	0.279	0.214
Full PCA*	0.240	0.281	0.212

Table 6: Performance of factor and PCA-based strategies across training and out-of-sample periods over the U.S. Treasuries. The table reports cumulative returns, Sharpe ratios, and turnover-adjusted Sharpe ratios over the training sample (2005–2013), two evaluation subperiods (2014–2019 and 2019–2024), and the full out-of-sample period (2014–2024).

Figure 5: Top row: cumulative returns of the NSS portfolio and the benchmarks. Bottom row: portfolio weights. The allocation is nearly static and concentrated in the shortest maturities, indicating that the apparent Sharpe outperformance mainly reflects a persistent low-duration tilt rather than active timing

Over the full evaluation sample, Carry and Modified Carry perform broadly in line with the benchmark after accounting for transaction costs, while momentum-based strategies do not outperform in this setting. Although both signals exhibit clear timing value when applied to the global WGBI universe, their informational content weakens substantially when restricted to the U.S. Treasury market. Signal magnitudes compress, and the resulting allocations remain close to the benchmark, indicating limited duration-timing ability. By contrast, the IC approach emerges as the most economically stable. It delivers strong full-sample performance and reacts appropriately during broad sell-offs, achieving higher Sharpe ratios and cumulative returns.

In this setting, Logistic Regression performs better than in the WGBI setting on average, particularly when combined with PCA, and retains part of this advantage even after accounting for transaction costs. By contrast, strategies based purely on PCA perform less well than the equal-weighted benchmark, indicating that dimensionality reduction alone is not sufficient to extract meaningful duration-timing information.

Across strategies, the p-values (see Table 7) are generally high, confirming the limited statistical strength of most factor-based timing rules in the U.S. Treasury market. With the exception of NSS, none of the strategies approaches conventional rejection thresholds under either HAC or bootstrap inference. Even for NSS, the p-values should be interpreted cautiously given its nearly static low-duration tilt documented earlier.

This outcome is consistent with the higher informational efficiency of a single, deep market: much of the information exploited by carry and momentum at the global level is already reflected in U.S. Treasury prices, reducing their marginal predictive power.

As a result, the Signal Combined strategy also fails to deliver meaningful improvements. When individual factors are weak, aggregating them into a single composite signal is insufficient to extract additional structure from the U.S. yield curve.

Given the evidence presented in this section, the IC strategy remains the most defensible choice among the factor-based approaches from a practical perspective, even though its statistical support remains limited in absolute terms.

Nevertheless, these findings suggest that yield-curve information alone is insufficient to generate robust duration-timing value in the U.S. Treasury market. This motivates extending the analysis beyond term-structure signals by incorporating macroeconomic and market-based time-series information in the next section.

Strategy	p_{HAC}	p_{Boot}
NSS	0.071	0.057
Log Reg Regime	0.243	0.244
Value	0.372	0.347
IC	0.432	0.429
Carry	0.485	0.512
Log Reg	0.532	0.539
Carry Mod	0.563	0.573
Full PCA	0.645	0.652
Full PCA*	0.653	0.632
Momentum	0.653	0.643
Momentum4	0.705	0.714
Momentum2	0.754	0.712
Momentum3	0.762	0.770
Signal Combined	0.773	0.785

Table 7: HAC-based and bootstrap p-values for the full evaluation sample (2014–2024) on U.S. Treasuries, reported in ascending order with respect to the HAC-based p-values. The table reports one-sided tests of the null hypothesis that each strategy does not outperform the equal-weighted benchmark in risk-adjusted terms.

4.2 Incorporating Market and Macro Signals with Machine Learning

a. Target Construction and Learning Setup

A natural way to incorporate additional signals and features is through machine-learning methods. Since there is no unambiguous target, we frame the problem as a classification task and train models to predict the bucket with the highest return over the next 21 trading days.

At each rebalancing date t, the target is constructed using 21-day forward returns, computed up to $t - 21$, ensuring that no forward-looking information is used. Before selecting the models, we analyze the empirical properties of this target. Figure 6 reports the histogram of the best-performing buckets in the training sample, i.e. over the 2005–2013 period. We can clearly see that the shortest-duration and the longest-duration buckets dominate the distribution, while intermediate maturities are selected far less frequently.

This effect becomes even more pronounced once we exclude borderline observations, defined as cases in which the return difference between the best and second-best buckets over the 21-day horizon is smaller than 20 basis points. After applying this margin filter, the intermediate-duration buckets are almost never selected, indicating that economically meaningful signals tend to correspond to a clear preference for either very short or very long duration.

For this reason, rather than training a multi-class classifier over all buckets, we focus on binary classification between the lowest-duration and highest-duration buckets. This choice avoids severe class imbalance and reduces model complexity, which is crucial given the high noise and weak signal present in this setting.

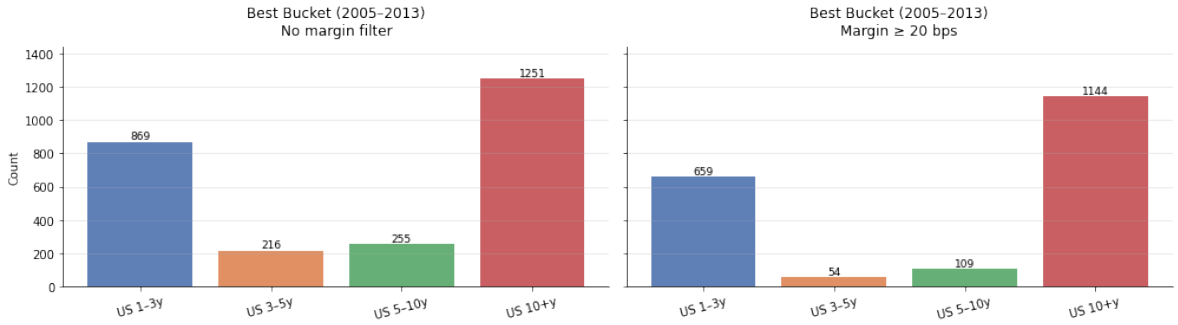

Figure 6: Histogram of the best-performing duration buckets based on 21-day forward returns in the training sample; excluding observations where the return spread between the best and second-best bucket is below 20 bps removes marginal cases and highlights that only the shortest- and longest-duration buckets remain dominant.

At each rebalancing date, models are retrained using a rolling window of the previous four years of daily data, corresponding to roughly 1,000 observations. This choice is motivated by regime considerations: using excessively long histories risks averaging over heterogeneous monetary and volatility regimes, while a four-year window is long enough to capture the prevailing macro-financial environment and short enough to adapt when regime shifts occur. The choice is also empirically robust, in the sense that nearby window lengths deliver very similar results and do not materially affect the conclusions. The raw observation count, however, is misleading. Financial time series exhibit strong serial dependence, and several predictors are observed only at a monthly frequency, so a substantial fraction of the data is redundant rather than independent. Once these effects are taken into account, the effective number of independent observations is much smaller (heuristically on the order of 50–100 data points per prediction), which constrains the complexity of models that can be reliably estimated.

In this setting, working with 15–20 time series and multiple transformations quickly leads to a feature space of roughly 75–100 variables, assuming that each time series is engineered into 5 features (on average). This places the problem in a regime in which even aggressive regularization is typically insufficient to prevent overfitting. For this reason, the only viable approach is to restrict attention to carefully selected subsets of predictors rather than using the full feature set at once; in practice, we limit each subset to at most ten features. This choice is a necessity in our framework and is consistent

with the empirical findings of [4], where they adopt a similar strategy to address the low effective dimensionality induced by correlated financial time series. While this approach may initially discard interactions across different subsets, we later attempt to recover and combine this information at a subsequent stage of the analysis.

Given these constraints, each model is trained to predict a single rebalancing decision, and the resulting probability outputs are directly used to construct portfolio weights. We use the following models:

- **Logistic Regression (LR).** Used as a simple and transparent benchmark. It is interpretable, and relatively robust in small-sample settings.
- **Random Forests (RF).** Included to capture nonlinear effects and interactions across features, at the cost of higher model complexity and an increased risk of overfitting.
- **Gradient Boosting (XGB).** Designed to model complex nonlinearities and feature interactions more aggressively than RF, but also more sensitive to overfitting in data-constrained environments.
- **Support Vector Machines (SVM) with Gaussian kernel.** Provide an intermediate level of flexibility, allowing for nonlinear decision boundaries while remaining more controlled than tree-based ensembles.
- **Stacked model.** Combines the probability outputs of RF, XGBoost, and SVM to aggregate their complementary strengths, with the caveat that stacking is intrinsically harder to regularize and more prone to overfitting given the limited effective sample size.

All hyperparameters are selected using the training sample only. In addition, models are estimated with exponentially decaying observation weights, so that more recent data receive higher importance. This choice is meant to improve adaptability to regime changes and allow the models to react more quickly to shifts in the underlying market environment.

b. Feature Engineering and Selection

As discussed, we partition the full information set into coherent subsets of predictors, based on similar economic interpretation. The predictor sets considered are:

- **Factor**: yield-curve and return-based signals capturing carry, value, momentum, and term-structure information (built as we did in the previous section).
- **Forex CHF**: USD–CHF exchange-rate.
- **Forex EUR**: USD–EUR exchange rate.
- **Gold**: gold price.

- **Macro**: mostly low-frequency macroeconomic variables capturing inflation dynamics and real-activity conditions. Specifically, we include CPI and PPI[1], the Federal Funds target rate (DFEDTAR) [2], survey-based expected inflation at the 1-, 5-, and 10-year horizons (EXPINF1YR, EXPINF5YR, EXPINF10YR), and a market-implied one-year inflation expectation change[3].

- **Market**: broad market indicators including the NASDAQ 100 (NDX), S&P 500 (SPX), VIX index, gold price (XAU), and the S&P GSCI commodity index, capturing global risk appetite, equity-market conditions, volatility, and commodity-cycle dynamics.

- **Gold & Commodities (Gold&Comm)**: gold price and the S&P GSCI commodity index, capturing safe-haven demand, inflation-hedging properties, global growth conditions, and commodity-cycle dynamics.

- **VIX**: equity-market volatility measure capturing changes in market uncertainty.

- **MOVE**: bond-market volatility index capturing uncertainty in interest-rate markets.

- **PCA**: principal components of the yield curve providing a low-dimensional representation of term-structure movements.

Each subset is then expanded through feature engineering, including trend and volatility measures, anomaly indicators, and simple interactions. Feature selection is carried out in two steps:

- **Correlation filtering**: highly collinear variables (within a subset) are first removed by excluding features with pairwise correlations above 0.90 or 0.95 (depending on the number of features obtained after engineering). See for example Figure 7, where this procedure is applied to the Macro subset.

- **Random Forest screening**: on the reduced feature set, a Random Forest is trained on realized future returns over the training sample to choose between long- and short-duration buckets. At this stage, the Random Forest's role is limited to identifying the most informative predictors within each subset, based on feature-importance scores. Since this screening step is performed exclusively on the training sample, and the selected features are subsequently fed into separate out-of-sample models, no forward-looking information is introduced. See for example Figure 8, where this procedure is applied to the Macro subset.

[1]Consumer Price Index and Producer Price Index. For each month, both series are aligned to the release date of the later of the two announcements (relative to the same month).

[2]We use the effective target (DFEDTAR) when available (up to 2008), and thereafter proxy it by the midpoint of the upper and lower bounds (DFEDTARU and DFEDTARL).

[3]It is constructed as the difference between the 1-year nominal Treasury yield (DGS1) and the yield on a 1-year inflation-linked zero-coupon bond.

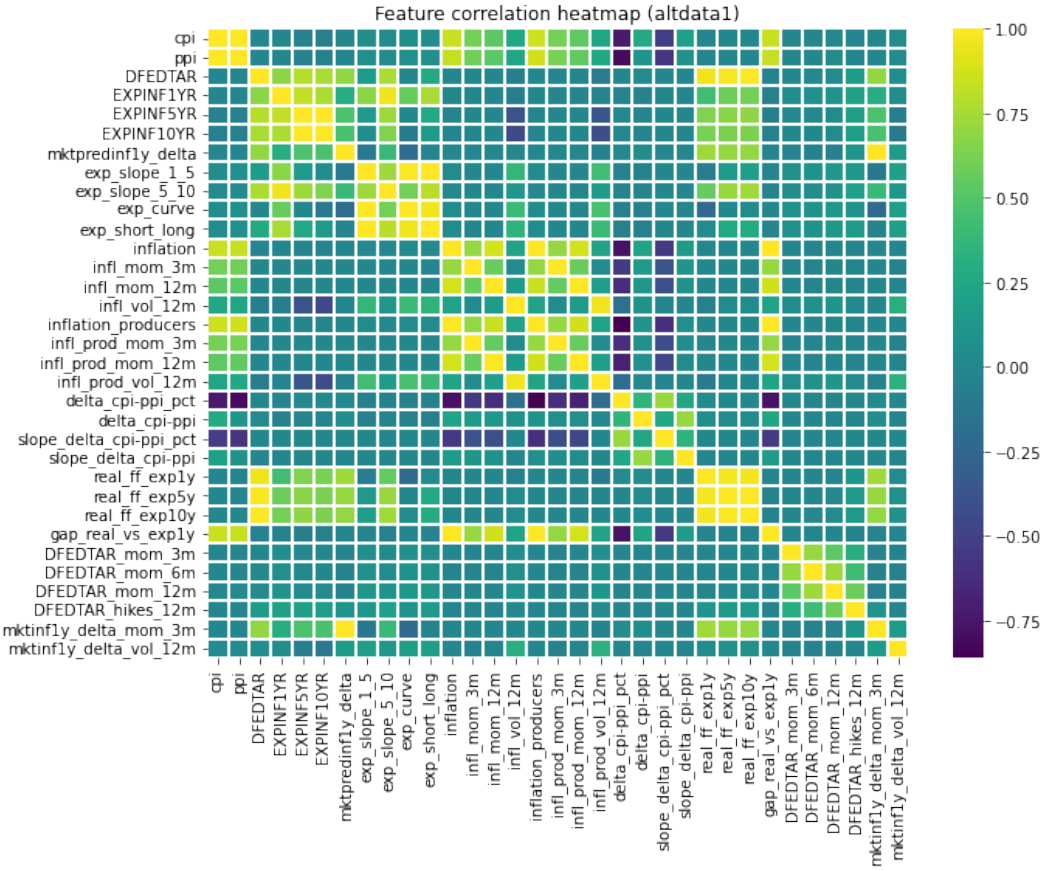

Figure 7: Correlation heatmap of the engineered features from the subset Macro over the 2005-2013 period. Strong dependencies across variables motivate a correlation-based pruning step prior to model training.

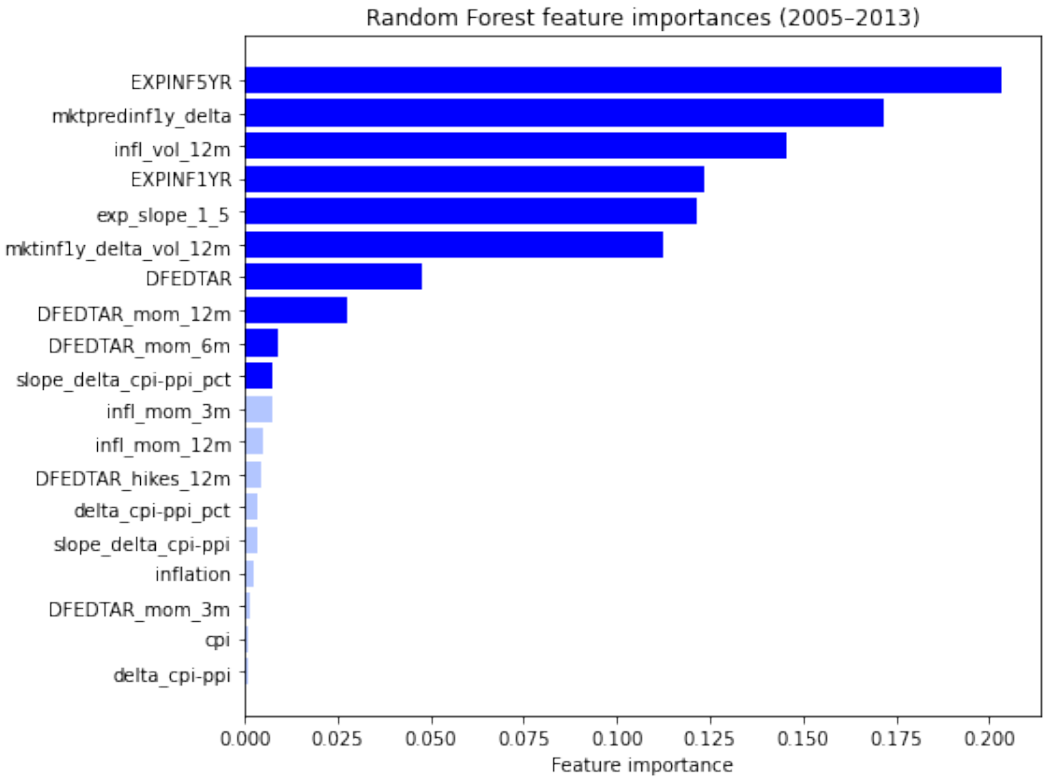

Figure 8: Random Forest feature importances estimated on the training sample (2005–2013) using only the subset of non-collinear features obtained from the subset Macro. Dark blue bars correspond to the features selected for subsequent modeling.

c. Results and Comments

Here we turn to the out-of-sample evaluation and analyze how the different strategies perform over the full evaluation period. Table 8 reports cumulative returns, Sharpe ratios and p-values computed over 2014–2024, allowing a direct comparison across models and information sets under identical market conditions. The methodology used to obtain the p-values have been described in in Section 1.1. For computational efficiency, results are reported only over the full evaluation window rather than split into subperiods: model training is repeated at each rebalancing date, which makes a finer temporal decomposition prohibitively costly. Moreover, performance on the training sample is of secondary interest at this stage, as feature selection has already been carried out using the training data and the focus here is on assessing out-of-sample generalization.

Statistical significance is reported in this setting because the strategies rely on alternative datasets applied to the U.S. bond market, where the economic value of additional information is not guaranteed ex ante. When models are enriched with macroeconomic, financial, or cross-asset signals, improvements in Sharpe ratios may reflect genuine informational content or simply increased model flexibility. Reporting p-values allows us to assess whether the apparent performance gains relative to the equal-weighted benchmark are statistically consistent with the presence of a true predictive signal in this specific market environment.

Full Evaluation Sample (2014–2024)

Strategy	Dataset	Cum. Return	Sharpe	Sharpe/turn	p_{HAC}	p_{Boot}
XGB	Forex CHF	0.505	0.479	0.432	0.201	0.178
RF	Forex CHF	0.410	0.457	0.419	0.152	0.130
LR	Factor	0.417	0.405	0.376	0.271	0.262
SVM	Forex CHF	0.363	0.405	0.359	0.342	0.349
SVM	MOVE	0.381	0.386	0.354	0.322	0.299
SVM	Macro	0.349	0.376	0.341	0.344	0.316
Stacked	Gold&Comm	0.314	0.347	0.331	0.319	0.300
SVM	Factor	0.328	0.363	0.321	0.418	0.393
Stacked	Macro	0.278	0.321	0.304	0.428	0.423
LR	Macro	0.285	0.311	0.291	0.490	0.525
Stacked	Market	0.264	0.305	0.291	0.491	0.464
Equal-weighted	/	0.182	0.291	0.290	/	/
Stacked	Gold	0.245	0.298	0.284	0.528	0.521
LR	Forex CHF	0.232	0.301	0.277	0.543	0.542
XGB	Forex EUR	0.282	0.320	0.273	0.536	0.550
XGB	MOVE	0.274	0.306	0.265	0.563	0.561
Stacked	Factor	0.232	0.271	0.254	0.646	0.644
Stacked	Forex CHF	0.204	0.252	0.236	0.744	0.739
Stacked	VIX	0.200	0.245	0.232	0.774	0.782
SVM	Gold	0.232	0.280	0.231	0.656	0.640
RF	MOVE	0.214	0.262	0.228	0.680	0.669
Stacked	PCA	0.190	0.240	0.225	0.783	0.811
RF	VIX	0.194	0.260	0.224	0.679	0.668
SVM	VIX	0.214	0.267	0.223	0.670	0.671
Stacked	Forex EUR	0.187	0.240	0.223	0.759	0.770
SVM	PCA	0.228	0.267	0.222	0.670	0.687
RF	Factor	0.198	0.252	0.219	0.717	0.707
RF	Macro	0.215	0.249	0.212	0.713	0.732
XGB	Macro	0.223	0.251	0.205	0.694	0.701
LR	VIX	0.172	0.232	0.205	0.741	0.722
Stacked	MOVE	0.153	0.199	0.185	0.907	0.911
LR	MOVE	0.151	0.209	0.185	0.795	0.791
LR	Gold	0.137	0.196	0.173	0.852	0.854
XGB	Factor	0.162	0.203	0.165	0.785	0.796
RF	Forex EUR	0.146	0.198	0.161	0.843	0.836
LR	Forex EUR	0.130	0.181	0.160	0.858	0.869
RF	Market	0.138	0.193	0.158	0.858	0.871
RF	Gold	0.124	0.175	0.141	0.889	0.891
RF	PCA	0.117	0.164	0.137	0.926	0.912
XGB	VIX	0.125	0.179	0.134	0.821	0.823
XGB	Market	0.125	0.172	0.124	0.848	0.850
SVM	Forex EUR	0.124	0.170	0.114	0.875	0.878
LR	PCA	0.091	0.136	0.109	0.939	0.957
LR	Market	0.081	0.128	0.100	0.929	0.943
XGB	Gold&Comm	0.096	0.150	0.095	0.870	0.866
SVM	Gold&Comm	0.074	0.127	0.072	0.899	0.883
RF	Gold&Comm	0.055	0.102	0.067	0.961	0.959
SVM	Market	0.053	0.098	0.042	0.923	0.923
LR	Gold&Comm	0.006	0.044	0.012	0.981	0.983
XGB	PCA	0.002	0.045	0.007	0.986	0.991

Continued on next page

Strategy	Dataset	Cum. Return	Sharpe	Sharpe/turn	p-value	p-value (boot)
XGB	Gold	−0.001	0.041	0.000	0.958	0.966

Table 8: Performance comparison across datasets and models, sorted by turnover-adjusted Sharpe ratio. Row colors: green = Forex CHF, violet = Factor, yellow = Macro, blue = equal-weighted baseline.

The main insights derived from Table 8 are summarised below:

- **Performance relative to the benchmark.** Several algorithm–dataset combinations deliver higher cumulative returns and Sharpe ratios than the equal-weighted portfolio, suggesting that duration timing may benefit from conditioning on additional information even after accounting for transaction costs. However, the associated p-values indicate that statistical significance is generally weak over the full sample. In particular, while some strategies exhibit economically meaningful improvements in Sharpe ratios, the null hypothesis of equal risk-adjusted performance relative to the benchmark cannot be rejected at conventional significance levels. Importantly, the p-values obtained from the asymptotic HAC-based inference and those from the block bootstrap are broadly similar, reinforcing the internal coherence of the statistical evidence. This pattern suggests that, although certain information sets appear promising from an economic perspective, the evidence remains limited in statistical terms, possibly also reflecting the relatively small number of rebalancing observations available in the evaluation window.

- **Multiple testing considerations.** Applying multiple-testing corrections such as Bonferroni or Benjamini–Hochberg would mechanically push most adjusted p-values close to one in this setting, eliminating virtually all statistical rejections. Given the number of specifications and the moderate individual p-values, this outcome is expected. However, the analysis does not rely on a large-scale search over randomly generated strategies. The specifications are economically motivated and structured along predefined information sets and model classes. For this reason, we report unadjusted p-values as measures of statistical reliability for each comparison, while acknowledging that formal corrections would further weaken statistical significance. One might argue that, given the large number of specifications tested, some strategies outperform the benchmark simply by chance. The evidence, however, does not appear entirely consistent with a purely random explanation. If results were purely random, we would expect a more symmetric dispersion across information sets, instead certain information sets tend to rank systematically higher across algorithms, while others consistently underperform. Although statistical significance remains limited, this cross-sectional structure suggests that differences in informational content may play a role, rather than outcomes being driven exclusively by random variation.

- **Limited turnover costs.** Transaction costs remain limited across specifications. First, portfolio weights adjust smoothly over time because the training window at each rebalancing date largely overlaps with the previous one, implying that successive model estimates (and thus portfolio allocations) do not change abruptly. For

this reason, no additional ad-hoc smoothing of the signals is required. Second, the investment universe consists of only two duration buckets, which mechanically limits turnover. Moreover, predicted class probabilities are rarely extreme, so portfolio weights tend to remain close to interior values rather than switching aggressively between assets, further containing trading costs.

- **CHF, EUR and Gold signals.** Datasets based on USD/CHF deliver consistently strong performance across different algorithms, indicating that movements in the USD/CHF exchange rate contain predictive information for U.S. duration timing that is largely robust to the choice of learning method. This contrasts with both EUR-based and gold-based signals. The EUR exchange rate is more directly exposed to euro-area macroeconomic conditions and monetary policy, which tends to anchor EUR-based predictors to regional dynamics and limits their relevance for U.S. Treasuries. Gold-based signals, by contrast, appear to behave primarily as contemporaneous hedges rather than forward-looking predictors. As documented by Baur and McDermott [3], gold provides protection during periods of extreme market stress, but this effect typically materializes once adverse conditions are already present rather than in advance. Consistent with this evidence, gold-based predictors deliver limited timing value in our setting. Overall, CHF-based signals exhibit the most stable out-of-sample performance, which we interpret as an empirical regularity over the evaluation period rather than evidence of a structural advantage. These results suggest that, within this specific empirical setting, USD/CHF behaves as an effective proxy for global risk sentiment relevant to U.S. duration timing. This interpretation is consistent with the role of CHF as a traditional safe-haven currency, but we emphasize that the result is conditional on the sample, market, and strategy design, and should not be interpreted as a structural property.

- **Overfitting risk in Factor and Macro datasets.** Factor- and Macro-based signals appear more prone to overfitting. Their strongest results are obtained with simpler or more controlled models (e.g. Logistic Regression and SVM), while more complex approaches (Random Forests, Gradient Boosting, stacking) tend to perform poorly, consistent with the limited effective sample size. Nevertheless, both Factor and Macro datasets consistently show predictive power and tend to function as intended: they capture economically meaningful information and deliver robust signals when model complexity is kept under control, confirming their usefulness as core inputs.

- **Ambiguous role of MOVE.** The MOVE dataset exhibits mixed predictive performance. It performs relatively well when paired with SVM and Gradient Boosting, but delivers weak or inconsistent results under Logistic Regression and Random Forests. Overall, its predictive content appears fragile and model-dependent, making its economic role less clear than that of Forex or factor-based signals.

- **Weak performance of Market, VIX, and Commodities.** Market-wide indicators, volatility measures (VIX), and commodity-based datasets exhibit rela-

tively weak performance. This suggests limited incremental information for duration timing, despite their traditional role as risk indicators.

- **Sharpe ratios and economic significance.** Higher Sharpe ratios generally coincide with higher cumulative returns. This indicates that performance improvements are not driven solely by volatility compression but correspond to economically meaningful excess returns.

Overall, when paired with appropriate information sets and sufficiently constrained algorithms, machine learning methods exhibit economically meaningful improvements relative to relying exclusively on handcrafted factor signals. In particular, USD/CHF Forex, Factor, and Macro datasets tend to deliver the strongest and most consistent performance across models, while other datasets provide limited incremental contribution. Model complexity plays a central role: moderate non-linear flexibility (e.g. SVMs) often improves results, whereas more complex ensemble methods do not systematically outperform simpler specifications. Logistic Regression remains competitive, likely because a substantial portion of relevant interactions is already captured through feature engineering.

Predictive performance nevertheless appears state-dependent. The relevance of a given dataset varies across macroeconomic regimes, and some information sets seem to perform primarily during stress episodes while underperforming simpler factor-based approaches in more stable environments. Given that rolling ML models may adjust only gradually to regime shifts, we next consider threshold-based strategies that condition directly on the current level of selected indicators rather than on rolling predictive relationships. The following section introduces these threshold strategies and subsequently proposes a systematic procedure for selecting datasets and algorithms based on recent out-of-sample performance (see Section 4.4).

4.3 Incorporating Market and Macro Signals with Threshold-Based Strategies

a. Threshold-Based Strategies

These strategies are deliberately simple and rule-based, with the aim of limiting overfitting and preserving economic interpretability. Given the limited sample size, more complex specifications cannot be evaluated reliably; threshold-based rules therefore provide a transparent and robust way to incorporate Market and Macro information without relying on in-sample noise. They can be interpreted as low-complexity counterparts to the machine-learning approaches discussed earlier, and serve as explicit, interpretable benchmarks.

This simplicity comes with clear limitations. First, the strategies perform poorly in benign, falling-yield environments, where carry dominates and defensive positioning is penalized, as observed in the training sample. This behavior is structural rather than accidental: the rules are designed for regime identification and downside protection, not for harvesting carry. Second, thresholds are calibrated once on the training sample

and kept fixed thereafter, introducing calibration risk as regime boundaries may shift over time.

As a result, these strategies are most effective during sharp sell-offs or periods of elevated stress, and tend to underperform in prolonged low-volatility regimes where simpler factor-based allocations remain more effective.

We designed and tested the following threshold-based strategies:

- **Fed-rate threshold strategy (Fed rate).**
  - Compute the 3-month change in the policy rate, Δr_{3m}.
  - If $\Delta r_{3m} > 0$: allocate 100% to 1–3y.
  - If $\Delta r_{3m} \leq -0.5$: allocate 50% to 3–5y and 50% to 5–10y.
  - Otherwise: allocate 100% to 10+y.

  Economic rationale: policy tightening raises short-end risk and favors short duration; strong easing supports intermediate maturities, while stable policy environments favor long duration through carry.

- **CPI–PPI sign strategy (CPI–PPI).**
  - Compute YoY inflation rates CPI_{YoY} and PPI_{YoY}.
  - If $\text{sign}(\text{CPI}_{\text{YoY}}) = \text{sign}(\text{PPI}_{\text{YoY}})$: allocate 100% to 10+y.
  - Otherwise: allocate 100% to 1–3y.

  Economic rationale: consistent CPI and PPI signals indicate coherent inflation dynamics, supporting long duration; divergence increases inflation uncertainty and favors short duration.

- **CPI level–momentum threshold strategy (CPI Lvl–Mom).**
  - Let π_t denote CPI YoY inflation and $\overline{\pi}_t^{(3)}$ its 3-month moving average.
  - Define momentum $z_t = \pi_t - \overline{\pi}_t^{(3)}$.
  - If $\pi_t \leq 2.0$ and $z_t \leq 0$: allocate 100% to 10+y.
  - If $\pi_t \geq 3.0$ and $z_t \geq 0$: allocate 100% to 1–3y.
  - Otherwise: allocate 50% to 3–5y and 50% to 5–10y.

  Economic rationale: low and decelerating inflation supports long duration, while high and accelerating inflation increases tightening risk and favors short duration.

- **Forex CHF z-score threshold strategy (Forex CHF).**
  - Define USD/CHF returns $r_t = \Delta(\text{USDCHF})_t$ and CHF strength

$$s_t = -\sum_{j=1}^{20} r_{t-j}.$$

- Standardize using a 252-day rolling window to obtain a lagged z-score z_t.
- If $z_t \geq 0.5$: allocate 100% to 10+y.
- If $z_t \leq -0.5$: allocate 100% to 1–3y.
- Otherwise: allocate 50% to 3–5y and 50% to 5–10y.

Economic rationale: CHF strength signals risk-off conditions and declining yields, favoring long duration; CHF weakness reflects risk-on regimes and supports shorter duration.

- **VIX level threshold strategy (VIX).**
  - Let VIX_t denote equity-market volatility.
  - If $\text{VIX}_t \leq 18$: allocate 100% to 10+y.
  - If $\text{VIX}_t \geq 25$: allocate 100% to 1–3y.
  - Otherwise: allocate 50% to 3–5y and 50% to 5–10y.

  Economic rationale: low volatility reflects stable financial conditions, supporting long duration; elevated volatility signals stress and favors short duration.

- **MOVE level threshold strategy (MOVE).**
  - Let MOVE_t denote bond-market volatility.
  - If $\text{MOVE}_t \leq 70$: allocate 100% to 10+y.
  - If $\text{MOVE}_t \geq 100$: allocate 100% to 1–3y.
  - Otherwise: allocate 50% to 3–5y and 50% to 5–10y.

  Economic rationale: low rate volatility supports long duration, while high uncertainty about interest rates favors short duration.

- **MOVE level + spike overlay strategy (MOVE+Spike).**
  - Apply the MOVE level rule above.
  - Define a k-day change $\Delta_{10}\text{MOVE}_t = \text{MOVE}_t - \text{MOVE}_{t-10}$.
  - If $\Delta_{10}\text{MOVE}_t \geq 10$: override and allocate 100% to 1–3y.

  Economic rationale: sudden jumps in bond-market volatility signal abrupt repricing of rate risk and justify an immediate defensive shift to short duration.

- **PCA (PC2/PC3) threshold strategy (PCA).**
  - Use lagged yield-curve principal components $(\text{PC2}_{t-1}, \text{PC3}_{t-1})$.
  - If $\text{PC3}_{t-1} > 0.2$: allocate to the belly of the curve (3–5y and 5–10y).
  - If $\text{PC2}_{t-1} < -0.2$: allocate to the short end (1–3y).
  - Otherwise: allocate mostly to long duration (10+y).

Economic rationale: PC2 and PC3 capture slope and curvature regimes of the yield curve, guiding allocation across short, intermediate, and long maturities.

- **Expected-inflation slope threshold strategy (Exp. Infl.).**
  - Define $d_t = \text{EXPINF1YR}_t - \text{EXPINF10YR}_t$.
  - If $d_t > 0$: allocate 100% to 1–3y.
  - If $d_t < -0.3$: allocate 100% to 10+y.
  - Otherwise: allocate 50% to 3–5y and 50% to 5–10y.

  Economic rationale: front-loaded inflation expectations increase tightening risk and favor short duration, while declining short-term expectations support long duration.

b. Results and Comments

The results in Table 9 highlight a clear asymmetry across regimes. In the training sample (2005–2013), several threshold strategies generate high cumulative returns but generally lower Sharpe ratios than the equal-weighted benchmark, indicating that gains come at the cost of higher volatility. This suggests that these rules are not designed to dominate in smooth, carry-driven environments.

In Evaluation Period I (2014–2019), characterized by declining yields and low volatility, the equal-weighted portfolio remains a strong benchmark. Only a limited subset of strategies, most notably MOVE+Spike and CPI–PPI, outperform on both returns and Sharpe ratios. By contrast, during Evaluation Period II (2019–2024), marked by inflation shocks and sharp repricing, the pattern reverses. Volatility- and inflation-based rules such as MOVE+Spike, MOVE, Expected Inflation, Fed-rate, and VIX clearly outperform the benchmark, confirming their effectiveness in stress regimes.

Over the full evaluation period (2014–2024), this regime dependence persists: threshold strategies appear complementary rather than substitutive to factor timing, underperforming in stable markets but adding value during turbulent phases. Transaction costs remain contained throughout, as evidenced by the limited gap between gross and turnover-adjusted Sharpe ratios.

From a statistical perspective (see Table 10), the comparisons of the Sharpe ratio against the equal-weighted benchmark provide limited evidence of outperformance in the full sample. This is not surprising: in Evaluation Period I, most threshold strategies perform similarly to the benchmark, which mechanically weakens full-sample statistical rejection even if performance is stronger in Evaluation Period II. Although the second subperiod displays clearer economic gains, testing significance exclusively over 2019–2024 would rely on roughly 60 rebalancing observations, which is insufficient to obtain reliable inference. The moderate full-sample p-values therefore reflect both regime heterogeneity and the limited time-series dimension available for testing. Overall, the evidence suggests economically meaningful, regime-dependent improvements, while statistical power remains constrained.

Training Sample (2005–2013)

Strategy	Cum.	Sharpe	Sharpe/turn
U.S. 1–3y	0.281	1.813	1.813
U.S. 3–5y	0.474	1.126	1.126
U.S. 5–10y	0.667	0.873	0.873
U.S. 10+y	1.077	0.664	0.664
Equal-weighted	0.612	0.900	0.900
MOVE+Spike	0.313	0.526	0.476
CPI–PPI	0.867	0.653	0.639
Fed-rate	0.575	0.502	0.494
VIX	0.283	0.413	0.378
CPI Lvl–Mom	0.650	0.512	0.512
MOVE	0.209	0.401	0.358
Exp. Infl.	0.118	0.217	0.161
PCA	0.299	0.417	0.341
Forex CHF	0.286	0.430	0.305

Evaluation Period I (2014–2019)

Strategy	Cum.	Sharpe	Sharpe/turn
U.S. 1–3y	0.077	1.422	1.422
U.S. 3–5y	0.133	0.878	0.878
U.S. 5–10y	0.217	0.775	0.775
U.S. 10+y	0.542	0.683	0.683
Equal-weighted	0.236	0.791	0.791
MOVE+Spike	0.599	0.869	0.840
CPI–PPI	0.775	0.933	0.898
Fed-rate	0.460	0.660	0.630
VIX	0.466	0.678	0.644
CPI Lvl–Mom	0.542	0.683	0.683
MOVE	0.456	0.781	0.758
Exp. Infl.	0.277	0.691	0.622
PCA	0.234	0.541	0.462
Forex CHF	0.080	0.227	0.105

Evaluation Period II (2019–2024)

Strategy	Cum.	Sharpe	Sharpe/turn
U.S. 1–3y	0.110	1.023	1.023
U.S. 3–5y	0.078	0.347	0.347
U.S. 5–10y	0.039	0.131	0.131
U.S. 10+y	−0.121	−0.055	−0.055
Equal-weighted	0.030	0.106	0.106
MOVE+Spike	0.226	0.391	0.353
CPI–PPI	−0.377	−0.477	−0.500
Fed-rate	0.158	0.256	0.237
VIX	0.227	0.346	0.313
CPI Lvl–Mom	−0.121	−0.055	−0.056
MOVE	0.207	0.361	0.342
Exp. Infl.	0.176	0.714	0.669
PCA	−0.072	−0.052	−0.094
Forex CHF	0.010	0.061	−0.035

Full Evaluation Sample (2014–2024)

Strategy	Cum.	Sharpe	Sharpe/turn
U.S. 1–3y	0.154	0.960	0.960
U.S. 3–5y	0.161	0.437	0.437
U.S. 5–10y	0.177	0.294	0.294
U.S. 10+y	0.179	0.175	0.175
Equal-weighted	0.182	0.291	0.291
MOVE+Spike	0.570	0.509	0.467
CPI–PPI	0.656	0.407	0.383
Fed-rate	0.471	0.356	0.336
VIX	0.366	0.338	0.294
CPI Lvl–Mom	0.179	0.175	0.174
MOVE	0.445	0.449	0.420
Exp. Infl.	0.406	0.619	0.566
PCA	0.256	0.293	0.226
Forex CHF	0.106	0.143	0.059

Table 9: Performance of threshold-based duration-timing strategies across training and out-of-sample periods. The table reports cumulative returns, Sharpe ratios, and turnover-adjusted Sharpe ratios over the training sample (2005–2013), two evaluation subperiods (2014–2019 and 2019–2024), and the full out-of-sample period (2014–2024).

Strategy	p_{HAC}	p_{Boot}
Exp. Infl.	0.151	0.140
MOVE+Spike	0.210	0.219
CPI–PPI	0.215	0.207
MOVE	0.288	0.318
Fed-rate	0.397	0.396
VIX	0.492	0.510
PCA	0.630	0.625
Forex CHF	0.871	0.886
CPI Lvl–Mom	0.945	0.942

Table 10: One-sided p-values for Sharpe ratio differences relative to the equal-weighted benchmark over 2014–2024, computed using HAC-based inference and block bootstrap. Strategies are ordered by increasing HAC-based p-value.

4.4 An ensemble of All Strategies

a. Methodology

Ex ante, it is unclear which of the strategies presented so far will perform best in the next period. While this is inherently difficult, we rely on the empirical evidence in the literature suggesting that momentum can be informative. We therefore construct an ensemble approach that selects and combines strategies based on their recent performance. Concretely, at each rebalancing date we summarize recent performance using a single score, rank all strategies accordingly, and select the top three strategies. The portfolio weights implied by these strategies are then combined using weights proportional to their scores to form the allocation for the subsequent period. By construction, this aggregation may provide implicit protection against large drawdowns by diversifying across multiple signals rather than relying on a single strategy.

To construct this ranking, at each rebalancing date t, for each strategy i we compute its cumulative return C_i and Sharpe ratio S_i over a rolling evaluation window (of four years). Sharpe ratios are truncated below at zero to avoid rewarding strategies with poor risk-adjusted performance. We then define the normalized quantities

$$\text{cum\_norm}_i = \frac{C_i}{\sum_j C_j}, \qquad \text{Sharpe\_norm}_i = \frac{\max\{S_i, 0\}}{\sum_j \max\{S_j, 0\}},$$

and construct the overall score

$$\text{Score}_i = 0.7 \cdot \text{Sharpe\_norm}_i + 0.3 \cdot \text{cum\_norm}_i.$$

When all Sharpe ratios are non-positive, we set $\text{Sharpe\_norm}_i = 0$ for all i, and the score reduces to a purely return-based criterion. This score places primary emphasis on risk-adjusted performance while still accounting for absolute returns. Relying exclusively on Sharpe ratios would be misleading, as illustrated by the training sample: the shortest-duration buckets exhibit the highest Sharpe ratios but generate relatively small cumulative returns. Incorporating a return component therefore helps prevent systematically favoring strategies that remain persistently exposed to low-duration buckets.

To avoid any ex post selection on the evaluation set, we fix the metric as it is (i.e., the 70–30 weights) and always select the top three performing strategies. Also, the evaluation window is kept fixed. An ex post analysis, not discussed here, shows that the performance results presented are not heavily influenced by the choice of these parameters.

b. Results and Comments

In Table 11, we show the performance of the ensemble of strategies. As in the previous two subsections, we also report one-sided p-values for Sharpe ratio differences relative to the equal-weighted benchmark, computed using both HAC-based inference and block bootstrap. Factor refers to all single-signal factor strategies introduced earlier explicitly excluding both signal combinations and PCA-based strategies. ML includes all machine-learning strategies obtained from the combination of datasets and algorithms, excluding stacking methods. Threshold-based strategies constitute a separate class and include all rule-based allocations driven by predefined levels or regime indicators.

Consistent with the previous sections, we focus exclusively on the full evaluation sample, as evaluating the ensemble at intermediate frequencies would require re-estimating and re-ranking a large number of strategies and would substantially increase computational costs.

Full Evaluation Sample (2014-2024)

Strategy	Cum. Return	Sharpe	Sharpe/turn	p_{HAC}	p_{Boot}
Equal-weighted	0.182	0.291	0.290	/	/
Threshold	0.630	0.533	0.496	0.122	0.124
ML	0.198	0.242	0.210	0.756	0.754
Factor	0.093	0.163	0.112	0.897	0.888
Threshold + Factor	0.344	0.390	0.342	0.386	0.372
All	0.311	0.364	0.320	0.428	0.419

Table 11: Performance comparison across strategy groups relative to the equal-weighted benchmark, Full Evaluation Sample (2014–2024). p-values refer to one-sided tests of Sharpe ratio differences using HAC-based inference and block bootstrap.

At first glance, the aggregate *All* portfolio appears to perform better than the equal-weighted benchmark, suggesting that combining all strategy classes is effective (see Figure 9). A closer inspection, however, reveals that this performance is not driven by a balanced contribution across components. Instead, it is largely attributable to the threshold-based strategies (see Figure 10), which display substantially higher Sharpe ratios and cumulative returns over the evaluation period.

Machine-learning strategies (see Figure 11) contribute relatively little in this setting, likely because they do not adapt quickly enough to regime changes and therefore struggle when market conditions shift abruptly. Factor strategies also underperform within the ensemble, not because of a lack of signal, but rather due to the ensemble methodology itself: factor-based approaches such as IC perform well in isolation, indicating that aggregation methods exploiting the cross-sectional correlation structure of

signals are more suitable than simple performance-based selection. Overall, the results suggest that the apparent success of the full ensemble mainly reflects the effectiveness of threshold-based rules in stress and sell-off regimes, rather than uniformly strong performance across all strategy classes. More generally, simple momentum at the strategy level (i.e., selecting at each point in time the strategy that has performed best up to that date) does not appear to be effective in this setting, further suggesting that past relative performance alone is a weak guide to future allocation across heterogeneous strategy classes. Finally, turnover remains moderate across all ensemble specifications, and transaction costs do not materially erode performance, indicating that the reported results are implementable in practice and that trading costs are well under control, given the high liquidity of the underlying assets.

From a statistical perspective, the evidence is limited. None of the ensemble specifications achieves conventional 5% significance. The threshold-based strategies display the lowest p-values (around 12%), providing comparatively stronger, though still moderate, statistical support. By contrast, ML, Factor, Threshold+Factor, and the aggregate All portfolio yield substantially higher p-values, indicating no statistically robust outperformance relative to the benchmark over the evaluation window.

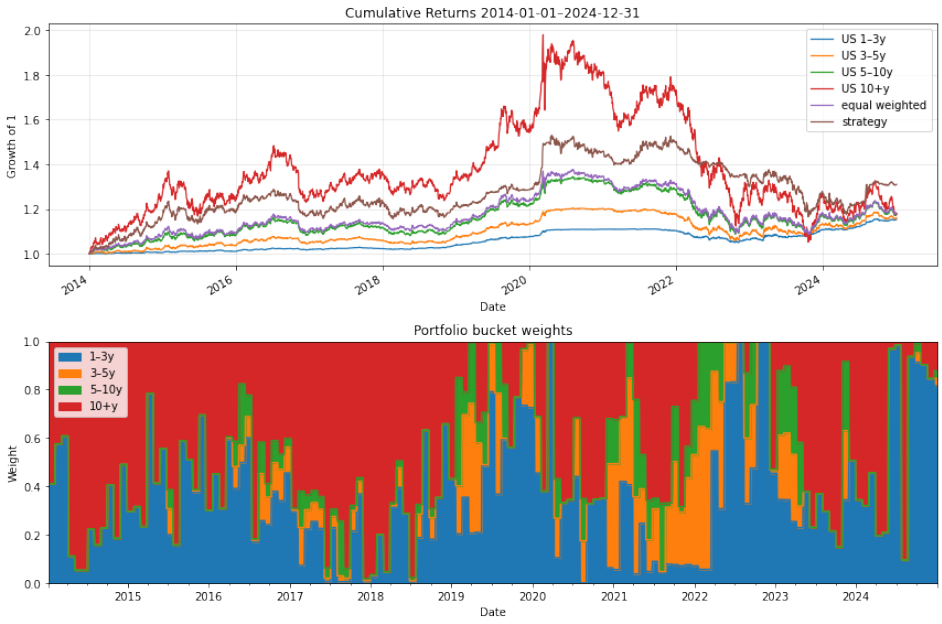

Figure 9: Full ensemble combining threshold-based, factor, and ML strategies. Top panel: cumulative returns compared with the benchmarks. Bottom panel: portfolio weights over time.

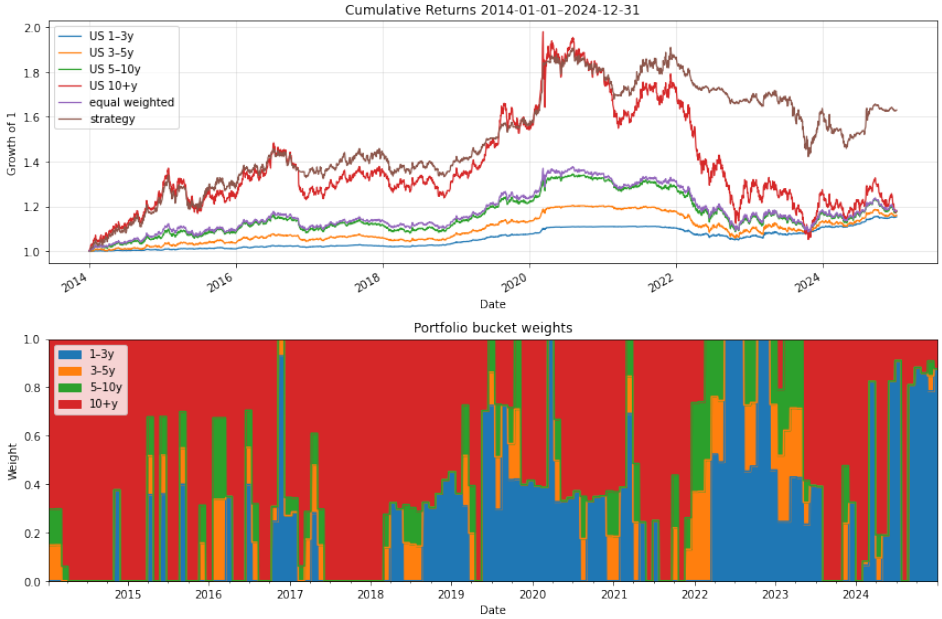

Figure 10: Threshold-based ensemble. Top panel: cumulative returns of the ensemble portfolio compared with the benchmarks. Bottom panel: portfolio weights over time.

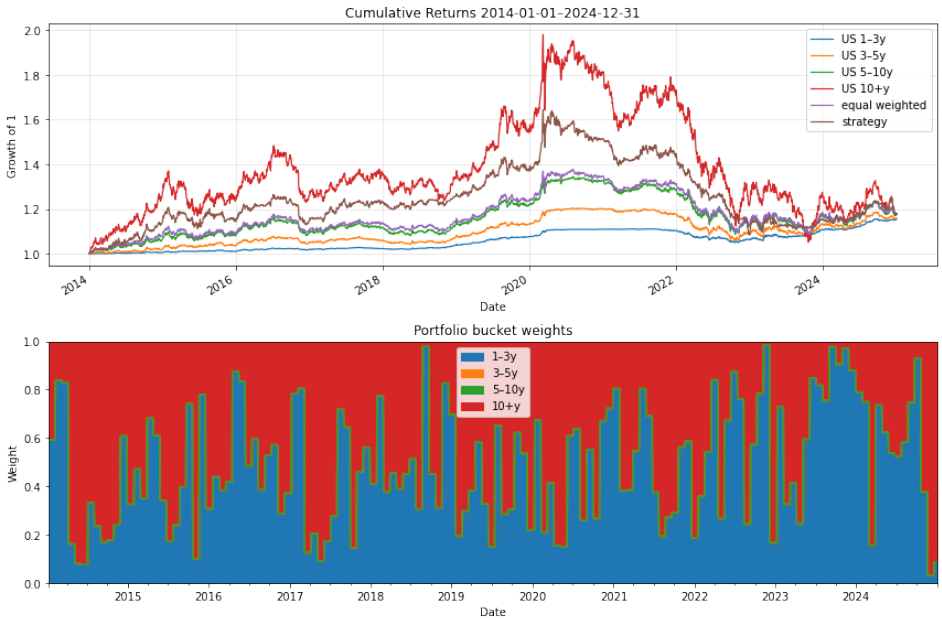

Figure 11: Machine-learning ensemble. Top panel: cumulative returns of the ML-based ensemble compared with the benchmarks. Bottom panel: portfolio weights over time.

5 IPCA Strategy on Developed Countries

In this third part, we extend the analysis to a multi-country, multi-index setting using Instrumented Principal Component Analysis (IPCA) to exploit cross-country variation in bond markets and identify common patterns across countries and maturities. To account for differences in market depth, transaction costs are set to 10 bps per unit of turnover, reflecting the lower liquidity of some constituent government bond markets.

5.1 IPCA model

IPCA provides a unified framework to model asset returns using latent risk factors whose loadings vary over time as a function of observable characteristics. In its canonical form, expected returns of asset i at time $t+1$, $r_{i,t+1}$, are written as

$$r_{i,t+1} = \beta_{i,t} f_{t+1} + \varepsilon_{i,t+1},$$

where $f_{t+1} \in \mathbb{R}^K$ is a vector of latent factors common across assets and $\beta_{i,t}$ denotes asset- and time-specific factor loadings. The key insight of IPCA is that these loadings are not free parameters but are instead modeled as linear functions of observable characteristics,

$$\beta_{i,t} = z_{i,t}^\top \Gamma,$$

where $z_{i,t} \in \mathbb{R}^L$ collects characteristics that vary across assets and over time, and $\Gamma \in \mathbb{R}^{L \times K}$ is a fixed mapping estimated from the data.

Estimation proceeds by jointly recovering the factor realizations $\{f_t\}_{t=1}^T$ and the characteristic-to-loading map Γ via an alternating least squares procedure. Specifically, the IPCA estimator solves the following least-squares problem:

$$\left(\hat{\Gamma}, \{\hat{f}_t\}_{t=1}^T\right) = \arg\min_{\Gamma, \{f_t\}_{t=1}^T} \sum_{t=1}^{T} \sum_{i=1}^{N} \left(r_{i,t+1} - z_{i,t}^\top \Gamma f_{t+1}\right)^2.$$

Details on statistical properties and estimation are provided in [14] and [15].

5.2 Returns Prediction and Portfolio Construction

The objective is again to select the most promising asset(s) from a broad cross-section of government bond indices. The investment universe consists of all four maturity buckets (1–3, 3–5, 5–10, and 10+ years) for 14 developed countries, yielding a total of $N = 56$ assets. Countries are included based on data availability, requiring that all four maturity buckets are consistently observable over the entire sample period. The final universe comprises Eurozone countries (Austria, Belgium, Germany, Spain, France, Italy, and the Netherlands) and non-Eurozone developed markets (Australia, Canada, Denmark, the United Kingdom, Japan, Sweden, and the United States). Predicted returns are subsequently translated into portfolio weights. We again use returns fully hedged into CHF to prevent FX movements from dominating the estimated common factors.

Portfolio decisions are updated every 21 business days. At each rebalancing date t, the IPCA model is estimated using cumulative 21-day returns over a rolling window

spanning from $t-T$ to $t-21$, where T denotes the length of the estimation window and is treated as a hyperparameter. This design ensures that only information available before portfolio formation is used, avoiding any forward-looking bias.

The estimation step produces an estimate of the characteristic-to-loading map $\widehat{\Gamma}$ together with a sequence of latent factor realizations $\{\widehat{f}_\tau\}_{\tau=t-T}^{t-21}$. Expected returns at time $t+21$ are then formed using lagged characteristics:

$$\widehat{r}_{i,t+21} = \widehat{\beta}_{i,t-1}\widehat{f}_t = z_{i,t-1}^\top \widehat{\Gamma}\,\widehat{f}_t,$$

where $z_{i,t-1}$ denotes the vector of observable characteristics for asset i available at the end of period $t-1$.

Since $\widehat{f}_t$ is not observed at the portfolio formation date, an additional forecasting step is required. We obtain $\widehat{f}_t$ by extrapolating the latent factor dynamics estimated over the training window. The specific forecasting rule for $\widehat{f}_t$, together with key hyperparameters such as the lookback window length T and the number of latent factors K, is selected empirically. We also assess whether fitting IPCA using log returns improves predictive performance relative to simple returns.

We begin by fixing the set of observable characteristics used in the IPCA model. These include an intercept, categorical indicators for maturity buckets (encoded using three dummy variables), and the three factor-based signals introduced earlier: carry (the average yield over the past five trading days), value (the deviation of the current yield from its five-year moving average), and momentum (the cumulative return over the previous month). This results in a total of $L = 7$ characteristics:

$$z_{i,t}^\top = \big[1,\ \mathbb{I}_{3\text{–}5,i},\ \mathbb{I}_{5\text{–}10,i},\ \mathbb{I}_{10+,i},\ \widetilde{s}_{i,t}^{\text{carry}},\ \widetilde{s}_{i,t}^{\text{value}},\ \widetilde{s}_{i,t}^{\text{momentum}}\big].$$

where $\widetilde{\cdot}$ denotes cross-sectional standardization across assets at each time t, and all characteristics are lagged so that only information available at t is used (avoiding look-ahead bias).

Because the value signal requires a five-year history of yields, the usable sample begins in 2010. Accordingly, the period 2010–2014 is used as an initial estimation window, and out-of-sample performance is evaluated from 2014 onward, consistent with the evaluation design adopted in the previous section.

We explore a range of modeling choices, considering both simple returns and log returns. Specifically, we vary:

- The number of latent factors: $K \in \{2, 3\}$.
- The length of the estimation window: $T \in \{3, 4, 5\}$ years.
- The method used to forecast latent factors:
  - Rolling averages over 1-, 3-, and $\max\{T, 5\}$-year windows.
  - Exponentially weighted moving averages (EWMA) with decay parameter $\lambda = 0.97$ over 1-, 3-, and $\max\{T, 5\}$-year windows.
  - A random-walk forecast using the last observed factor realization.
  - Autoregressive AR(1) forecasts estimated over 1-, 3-, and $\max\{T, 5\}$-year samples.

Model performance over the period 2010-2013 is evaluated using correlation-based measures: the Pearson correlation between predicted and realized returns and the Spearman rank correlation, which assesses the quality of cross-sectional rankings. These metrics are computed on a rolling basis every 10 days and subsequently averaged over time, increasing the effective evaluation sample size relative to using only monthly observations.

After computing the mean and standard deviation of these correlation measures (while noting that this summary ignores time-series dependence), we find that specifications based on log returns perform slightly better than those based on simple returns. For this reason, and because log returns are expected to offer greater numerical stability, we focus on log returns and report only results obtained using log returns.

As shown in Tables 12 and 13, a clear and consistent pattern emerges. Specifications with $K = 2$ latent factors systematically outperform those with $K = 3$, and longer estimation windows T are associated with improved performance. Moreover, the most effective forecasting approach is based on taking the time-series mean of the latent factors over a long horizon.

Accordingly, the selected specification employs $K = 2$ latent factors, is trained on a 4-year rolling window, and forecasts factor realizations by averaging over the entire estimation window. This configuration yields an average Information Coefficient (Pearson correlation) of 0.1430. A crude mean-to-std ratio (ignoring serial dependence) is $\frac{\overline{\text{IC}}}{\text{sd(IC)}} = \frac{0.1430}{0.4189} \approx 0.34$, indicating a small but economically meaningful predictive signal, which is encouraging given the low signal-to-noise ratio typically observed in government bond returns.

Table 12: Top 8 IPCA configurations ranked by average Pearson correlation over the 2010-2013 period

K	Train (y)	Avg (y)	Method	Pearson mean	Pearson med	Pearson std
2	4	4	Mean	0.1430	0.2312	0.4189
2	4	3	Mean	0.1315	0.2716	0.4426
2	3	3	Mean	0.1167	0.1904	0.4558
2	5	5	Mean	0.1140	0.1890	0.4414
2	5	3	Mean	0.1139	0.2053	0.4493
2	4	1	Mean	0.0558	0.0302	0.4609
2	4	-	Last	0.0555	0.0295	0.4612
2	4	5	AR(1)	0.0555	0.0295	0.4611

With this specification, we also consider an additional formulation in which categorical features are removed altogether and replaced by continuous bond characteristics: effective duration, convexity, and option-adjusted spread to swap (OAS). Effective duration measures the first-order sensitivity of bond prices to changes in yields while accounting for embedded optionality, convexity captures second-order non-linear price effects associated with yield movements, and the option-adjusted spread to swap reflects relative valuation and compensation for credit and liquidity risk net of option effects. As before, the number of features is capped at seven to remain consistent with

Table 13: Top 8 IPCA configurations ranked by average Spearman correlation over the 2010-2013 period

K	Train (y)	Avg (y)	Method	Spearman mean	Spearman med	Spearman std
2	4	4	Mean	0.1174	0.1606	0.3507
2	4	3	Mean	0.1113	0.1700	0.3798
2	3	3	Mean	0.1061	0.1298	0.3713
2	5	3	Mean	0.0999	0.1578	0.3744
2	5	5	Mean	0.0911	0.1398	0.3420
3	3	3	EWMA	0.0750	-0.0158	0.4240
3	3	1	EWMA	0.0747	-0.0158	0.4239
2	4	-	Last	0.0624	0.0441	0.4166

the cross-sectional dimension of 56 assets and to limit over-parameterisation:

$$z_{i,t}^{\top} = \big[1,\ \widetilde{\text{duration}}_{i,t},\ \widetilde{\text{convexity}}_{i,t},\ \widetilde{\text{OAS}}_{i,t},\ \widetilde{s}_{i,t}^{\text{carry}},\ \widetilde{s}_{i,t}^{\text{value}},\ \widetilde{s}_{i,t}^{\text{momentum}}\big].$$

Having obtained IPCA return forecasts, we translate them into portfolio weights using two allocation rules:

- **Softmax allocation.** Predicted returns are mapped into long-only portfolio weights (to be used at time t) via a softmax transformation with temperature parameter τ:

$$w_{i,t} = \frac{\exp\big(\widetilde{\widehat{r}}_{i,t+21}/\tau\big)}{\sum_{j=1}^{N}\exp\big(\widetilde{\widehat{r}}_{j,t+21}/\tau\big)}.$$

  Predicted returns are standardized cross-sectionally prior to applying the softmax. We consider $\tau \in \{0.5, 1.0\}$, where lower values of τ produce more concentrated portfolios.

- **Mean–variance allocation.** Since IPCA delivers direct forecasts of expected returns $\widehat{r}_{t+21} = (\widehat{r}_{1,t+21}, \ldots, \widehat{r}_{N,t+21})$, we also construct portfolios by solving a constrained mean–variance optimization problem (implemented using the Python library `CVXPY`):

$$\max_{w}\ \widehat{r}_{t+21}^{\top} w - \frac{\gamma}{2}\, w^{\top}\widehat{\Sigma}_t^{\text{shr}} w,$$

  subject to

$$w \geq 0, \qquad \mathbf{1}^{\top} w = 1, \qquad w_i \leq 0.10\ \ \forall i,$$

  where γ denotes the risk-aversion parameter and $\widehat{\Sigma}_t^{\text{shr}}$ is a shrinkage estimator of the return covariance matrix $\widehat{\Sigma}_t$ (computed over 21-day returns over the estimation window up to $t-21$):

$$\widehat{\Sigma}_t^{\text{shr}} = (1-\alpha)\widehat{\Sigma}_t + \alpha\frac{\text{tr}(\widehat{\Sigma}_t)}{N} I.$$

  Shrinkage is employed to mitigate estimation noise and improve numerical stability. We set $\alpha = 0.4$ and test $\gamma \in \{5, 10, 30\}$. Upper bounds on individual weights are imposed to enforce diversification.

5.3 Results and Comments

Main results are derived from Tables 14 and 16, which report the allocation and performance outcomes of the IPCA-based strategies under the two alternative feature specifications.

In this setting, we consider three equal-weighted benchmark portfolios: (i) an equal-weighted allocation across the four WGBI duration buckets (1–3y, 3–5y, 5–10y, 10+y); (ii) a global equal-weighted portfolio obtained by equally weighting all country–duration combinations across these four buckets, for a total of 56 assets; and (iii) an equal-weighted U.S. portfolio constructed over the same four maturity segments.

We additionally compute one-sided p-values to assess whether the IPCA-based strategies outperform each benchmark in terms of net Sharpe ratio on the full evaluation window. The p-values, computed as specified in Section 1.1 using both HAC-based inference and block bootstrap procedures, are reported in Tables 15 and 17.

The main findings are summarized below:

- **Benchmarks and IPCA performance.** Across all subperiods, IPCA-based allocation strategies consistently outperform these benchmarks in terms of cumulative returns. During the training sample, this return outperformance does not always translate into higher Sharpe ratios, as the equal-weighted portfolios already exhibit strong risk-adjusted performance. In this phase, IPCA-based strategies tend to allocate more aggressively toward longer-duration segments of the yield curve, generating higher returns at the cost of increased volatility. In contrast, during the out-of-sample evaluation periods, IPCA-based strategies achieve higher Sharpe ratios. This improvement is driven by their ability to dynamically reallocate exposure toward shorter-duration buckets in adverse market conditions, thereby limiting drawdowns and reducing volatility. As a result, the return profile becomes more balanced and risk-adjusted performance improves.

- **Statistical evidence relative to benchmarks.** Tables 15 and 17 report one-sided tests of Sharpe ratio differences between each IPCA-based strategy and the three equal-weighted benchmarks. Overall, statistical evidence is strongest relative to the U.S. and WGBI benchmarks, while results against the global All-country benchmark are generally weaker. Lower p-values are concentrated in a limited subset of specifications, notably those based on duration dummy variables combined with softmax allocation, with some approaching or falling below the conventional 5% significance level. Other parameterizations yield weaker statistical support. No multiple-testing correction is applied, as the specifications arise from economically structured design choices rather than from an unrestricted search over model space. The reported p-values should therefore be interpreted as indicators of the statistical reliability of each comparison rather than as evidence of systematic dominance. The following bullet points examine how differences in feature representation and portfolio construction drive these patterns in both economic and statistical performance.

- **Feature specification.** Comparing the two feature sets, the IPCA specification based on duration dummy variables performs better overall than the specification

based on continuous bond characteristics, with the advantage being particularly pronounced out of sample, especially during Evaluation Period I (2014–2019). This suggests that explicitly grouping bonds into maturity buckets and modelling comparable segments of the yield curve improves stability and reduces estimation noise. A likely explanation is the strong collinearity among continuous bond characteristics such as duration, convexity, and OAS, which can destabilize the estimation of the map Γ, whereas duration dummies provide a more stable discretization of the yield curve. This difference is also reflected in the p-values: specifications based on duration dummies generally achieve lower one-sided p-values across benchmarks, while continuous-characteristic models tend to deliver weaker statistical evidence.

- **Transaction costs.** Net Sharpe ratios are computed using transaction costs as high as 10 bps to reflect the lower liquidity of some constituent markets. Transaction costs play a non-negligible role, as IPCA-based strategies involve higher turnover. Nevertheless, even after accounting for these costs, IPCA-based portfolios continue to outperform the equal-weighted benchmarks in terms of net Sharpe ratios, confirming the robustness of the results. As explained in Section 1.1, all reported p-values are computed using net returns, i.e. after transaction costs, so that statistical inference reflects implementable performance.

- **Softmax versus mean–variance allocation.** Softmax and mean–variance allocation schemes deliver broadly similar performance in terms of gross returns and Sharpe ratios. However, softmax portfolios tend to produce more uniform weights (over time) and significantly lower turnover, resulting in materially lower transaction costs. In the mean–variance case, imposing an upper bound of 0.10 on individual weights enforces a more distributed allocation and prevents excessive concentration. Nevertheless, the optimization still leads to higher turnover than softmax, as weights react more strongly to changes in expected returns and covariance estimates. For this reason, softmax allocation appears preferable in practice, particularly in lower-liquidity bond markets. For visual evidence, see the second rows of figures 13 and 14. This difference is also reflected in the statistical results: softmax-based specifications generally achieve lower p-values across benchmarks (particularly under the duration-dummy specification), whereas mean–variance allocations tend to produce weaker statistical evidence of outperformance.

- **Role of allocation parameters.** Within the softmax framework, lower temperature values (corresponding to more confident and concentrated allocations) lead to higher cumulative returns at the cost of slightly higher volatility. In this setting, increased confidence in the signal proves beneficial in terms of pure Sharpe ratio, as the rise in risk remains limited. In terms of p-values, however, the different temperature specifications appear broadly equivalent, suggesting that the statistical strength of the signal is not highly sensitive to this parameter.

  A similar trade-off is observed for mean–variance portfolios: lower risk aversion (e.g., $\gamma = 5$) yields higher returns and, in this specification, also lower turnover (compared to mean–variance portfolios with $\gamma = 10$ and $\gamma = 30$). This behaviour may be related to instability in the estimated covariance matrix, which provides

further justification for the use of covariance shrinkage. Without shrinkage, these effects would likely be more pronounced, leading to excessively concentrated and unstable allocations and even higher turnover.

From a statistical perspective, lower values of γ tend to be associated with lower p-values, indicating comparatively stronger evidence of outperformance. As γ increases, both economic and statistical performance weaken, consistent with the progressively more conservative allocation implied by higher risk aversion.

- **Duration allocation versus asset selection.** The two allocation schemes differ in how they exploit the IPCA signal. Mean–variance portfolios primarily act as *duration selectors*: across all values of γ, the optimization concentrates risk along maturity buckets rather than on individual countries. In particular, during market stress, MV portfolios systematically shift exposure toward shorter-duration segments of the curve, reducing drawdowns regardless of the country composition. By contrast, softmax allocations exhibit a stronger *asset-selection* component. With a lower temperature ($\tau = 0.5$), the portfolio tends to concentrate on a small number of countries, selecting those with the strongest signals (notably Japan, which performs well over the sample). With a higher temperature ($\tau = 1.0$), the allocation becomes more diversified across assets sharing similar duration characteristics, emphasizing duration exposure while avoiding excessive country concentration. For visual evidence, see the third rows of figures 13 and 14.

Overall, the results suggest that the most implementable and statistically consistent specification is a softmax allocation with duration dummy variables and a moderate temperature (e.g. $\tau = 1$). This setup delivers strong out-of-sample Sharpe ratios, avoids excessive concentration on single assets, and keeps turnover and transaction costs materially lower than mean–variance portfolios. Although the mean–variance strategy (with dummy variables) and low risk aversion ($\gamma = 5$) shows greater resilience during severe drawdowns, it involves higher turnover and, in relative terms, higher p-values, making it less attractive from a joint economic–statistical perspective.

Tables for IPCA-based strategies using signals and duration dummy variables

Training Sample (2010–2013)

Strategy	Cum.	Sharpe	Sharpe/turn
EW (WGBI)	0.125	1.246	1.244
EW (Global)	0.157	1.231	1.227
EW (US)	0.156	0.676	0.674
SM ($\tau = 0.5$)	0.302	0.707	0.653
SM ($\tau = 1.0$)	0.254	0.952	0.901
MV ($\gamma = 5$)	0.182	0.639	0.578
MV ($\gamma = 10$)	0.160	0.592	0.525
MV ($\gamma = 30$)	0.137	0.597	0.513

Evaluation Period I (2014–2019)

Strategy	Cum.	Sharpe	Sharpe/turn
EW (WGBI)	0.121	0.784	0.781
EW (Global)	0.159	0.914	0.910
EW (US)	0.085	0.316	0.314
SM ($\tau = 0.5$)	0.318	1.065	0.976
SM ($\tau = 1.0$)	0.275	1.044	0.995
MV ($\gamma = 5$)	0.266	0.934	0.859
MV ($\gamma = 10$)	0.235	0.919	0.834
MV ($\gamma = 30$)	0.161	0.915	0.811

Evaluation Period II (2019–2024)

Strategy	Cum.	Sharpe	Sharpe/turn
EW (WGBI)	−0.102	−0.410	−0.411
EW (Global)	−0.109	−0.400	−0.403
EW (US)	−0.140	−0.346	−0.347
SM ($\tau = 0.5$)	−0.032	−0.089	−0.333
SM ($\tau = 1.0$)	−0.062	−0.214	−0.376
MV ($\gamma = 5$)	0.012	0.068	−0.193
MV ($\gamma = 10$)	−0.014	−0.045	−0.341
MV ($\gamma = 30$)	−0.035	−0.206	−0.581

Full Evaluation Sample (2014–2024)

Strategy	Cum.	Sharpe	Sharpe/turn
EW (WGBI)	−0.027	−0.055	−0.057
EW (Global)	−0.001	0.015	0.012
EW (US)	−0.104	−0.145	−0.147
SM ($\tau = 0.5$)	0.239	0.422	0.257
SM ($\tau = 1.0$)	0.132	0.280	0.173
MV ($\gamma = 5$)	0.154	0.319	0.159
MV ($\gamma = 10$)	0.125	0.294	0.112
MV ($\gamma = 30$)	0.062	0.207	−0.025

Table 14: IPCA-based allocation results using signals and duration dummy variables. EW denotes equal-weighted benchmarks, SM softmax allocation, and MV mean–variance portfolios.

Strategy	$p_{\text{HAC, WGBI}}$	$p_{\text{Boot, WGBI}}$	$p_{\text{HAC, All}}$	$p_{\text{Boot, All}}$	$p_{\text{HAC, US}}$	$p_{\text{Boot, US}}$
SM ($\tau = 0.5$)	0.061	0.056	0.120	0.139	0.027	0.027
SM ($\tau = 1.0$)	0.052	0.056	0.115	0.117	0.044	0.049
MV ($\gamma = 5$)	0.124	0.125	0.220	0.248	0.073	0.081
MV ($\gamma = 10$)	0.185	0.185	0.300	0.302	0.114	0.132
MV ($\gamma = 30$)	0.429	0.431	0.583	0.580	0.286	0.279

Table 15: One-sided p-values for net Sharpe ratio differences of IPCA-based strategies (using signals and duration dummy variables) relative to equal-weighted benchmarks (WGBI, All-country, and U.S.) over the full evaluation period (2014-2024). Inference is conducted using both HAC-based asymptotics and block bootstrap.

Tables for IPCA-based strategies using signals and continuous bond characteristics

Training Sample (2010–2013)

Strategy	Cum.	Sharpe	Sharpe/turn
EW (WGBI)	0.125	1.246	1.244
EW (Global)	0.157	1.231	1.227
EW (US)	0.156	0.676	0.674
SM ($\tau = 0.5$)	0.270	0.690	0.617
SM ($\tau = 1.0$)	0.226	0.945	0.875
MV ($\gamma = 5$)	0.191	0.724	0.639
MV ($\gamma = 10$)	0.168	0.687	0.596
MV ($\gamma = 30$)	0.143	0.773	0.666

Evaluation Period I (2014–2019)

Strategy	Cum.	Sharpe	Sharpe/turn
EW (WGBI)	0.121	0.784	0.781
EW (Global)	0.159	0.914	0.910
EW (US)	0.085	0.316	0.314
SM ($\tau = 0.5$)	0.392	0.756	0.640
SM ($\tau = 1.0$)	0.294	0.881	0.781
MV ($\gamma = 5$)	0.284	0.960	0.829
MV ($\gamma = 10$)	0.195	0.815	0.669
MV ($\gamma = 30$)	0.090	0.609	0.449

Evaluation Period II (2019–2024)

Strategy	Cum.	Sharpe	Sharpe/turn
EW (WGBI)	−0.102	−0.410	−0.411
EW (Global)	−0.109	−0.400	−0.403
EW (US)	−0.140	−0.346	−0.347
SM ($\tau = 0.5$)	−0.010	0.010	−0.106
SM ($\tau = 1.0$)	−0.050	−0.155	−0.271
MV ($\gamma = 5$)	−0.026	−0.101	−0.298
MV ($\gamma = 10$)	−0.026	−0.132	−0.378
MV ($\gamma = 30$)	−0.046	−0.408	−0.737

Full Evaluation Sample (2014–2024)

Strategy	Cum.	Sharpe	Sharpe/turn
EW (WGBI)	−0.027	−0.055	−0.057
EW (Global)	−0.001	0.015	0.012
EW (US)	−0.104	−0.145	−0.147
SM ($\tau = 0.5$)	0.223	0.290	0.175
SM ($\tau = 1.0$)	0.132	0.255	0.147
MV ($\gamma = 5$)	0.156	0.344	0.184
MV ($\gamma = 10$)	0.089	0.245	0.066
MV ($\gamma = 30$)	0.019	0.090	−0.126

Table 16: IPCA-based allocation results using signals and continuous bond characteristics. EW denotes equal-weighted benchmarks, SM softmax allocation, and MV mean–variance portfolios.

Strategy	$p_{\text{HAC, WGBI}}$	$p_{\text{Boot, WGBI}}$	$p_{\text{HAC, All}}$	$p_{\text{Boot, All}}$	$p_{\text{HAC, US}}$	$p_{\text{Boot, US}}$
SM ($\tau = 0.5$)	0.210	0.203	0.287	0.296	0.142	0.164
SM ($\tau = 1.0$)	0.161	0.158	0.258	0.257	0.107	0.116
MV ($\gamma = 5$)	0.114	0.112	0.203	0.209	0.073	0.075
MV ($\gamma = 10$)	0.271	0.301	0.399	0.382	0.182	0.200
MV ($\gamma = 30$)	0.635	0.605	0.745	0.752	0.466	0.442

Table 17: One-sided p-values for net Sharpe ratio differences of IPCA-based strategies (using signals and continuous bond characteristics) relative to alternative equal-weighted benchmarks (WGBI, All-country, and U.S.) over the full evaluation period (2014-2024). Inference is conducted using both HAC-based asymptotics and block bootstrap.

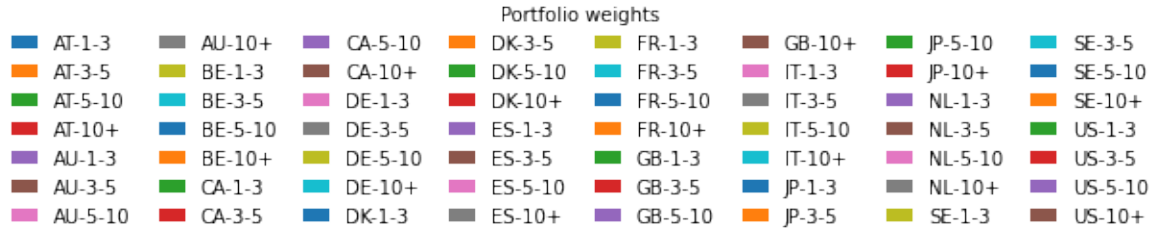

Figure 12: The legend refers to the second row of the panels below and should be read column by column. Assets are stacked in a fixed order: countries appearing first in the legend (e.g. Austria) are plotted at the bottom of the stacked area charts, with subsequent countries layered above.

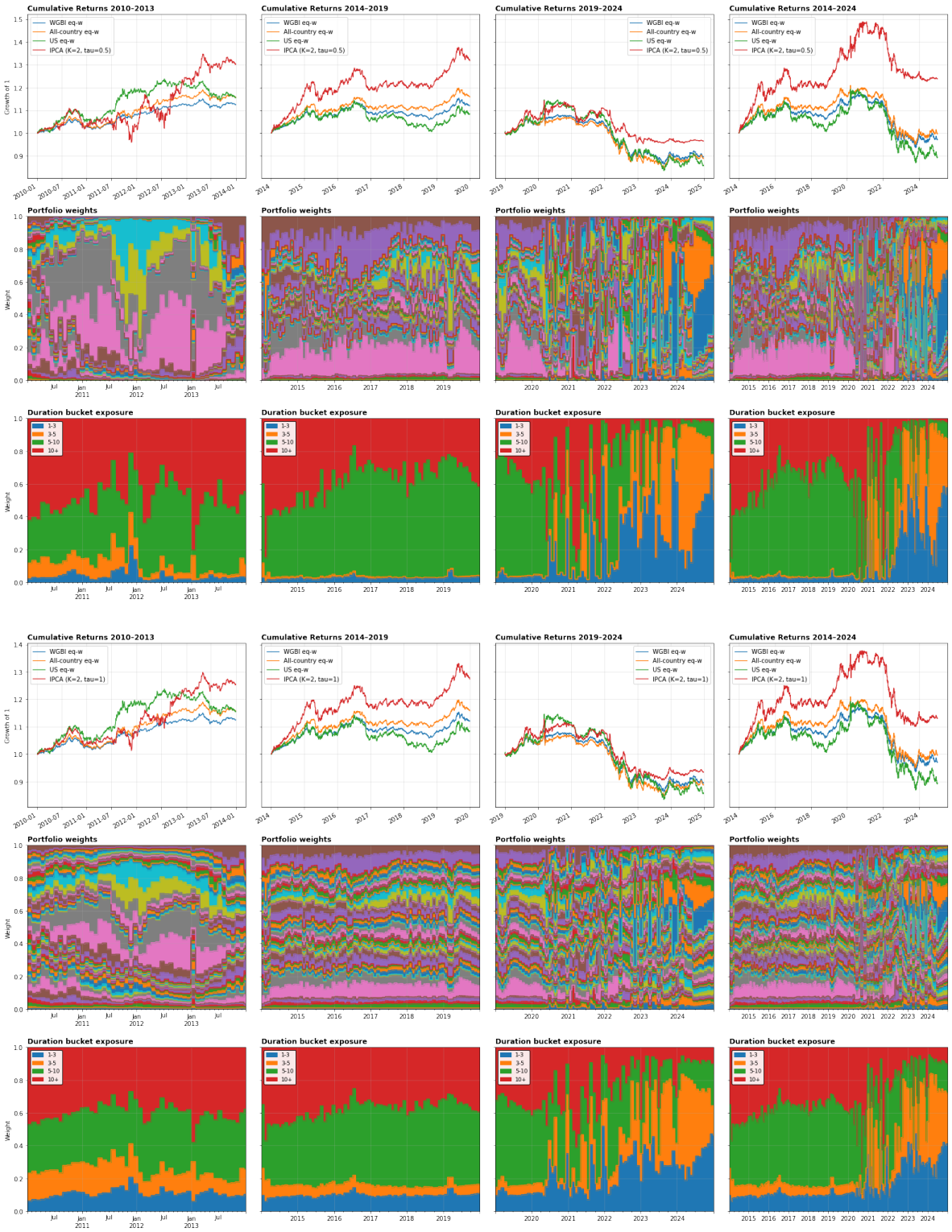

Figure 13: IPCA softmax allocation with $\tau = 0.5$ (upper panel), and $\tau = 1.0$ (lower panel) obtained using signals and dummy variables. The first row reports cumulative portfolio returns. The second row shows asset-level (country–maturity) weights: with $\tau = 0.5$ the allocation is more concentrated, especially toward the end of the sample (primarily on Japan), whereas $\tau = 1.0$ yields a more evenly distributed allocation. The third row reports weights aggregated by duration buckets, showing that both specifications shift exposure toward shorter durations during the drawdown period.

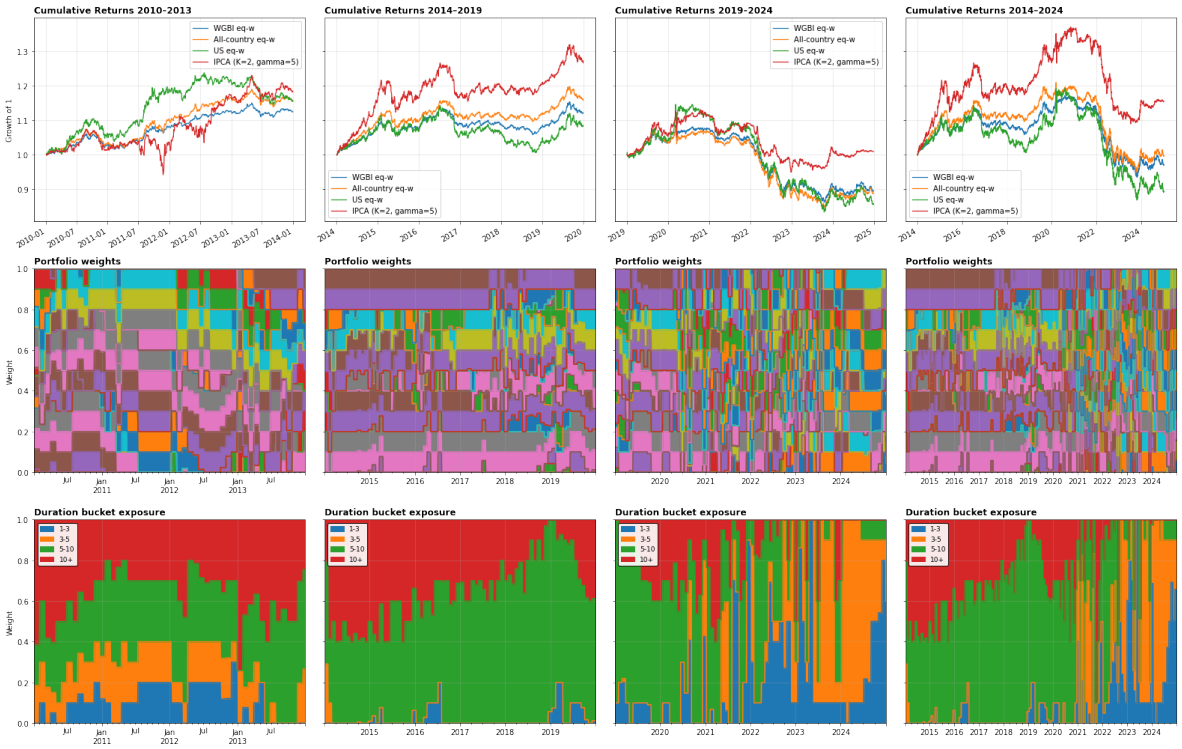

Figure 14: IPCA mean–variance allocation with risk-aversion parameter $\gamma = 5$. The first row reports cumulative portfolio returns. The second row shows asset-level (country–maturity) weights, highlighting that the weight cap at 0.10 is frequently binding during stable market conditions, while it is reached less often during drawdown periods. The third row reports weights aggregated by duration buckets, showing that the mean–variance allocation systematically shifts exposure toward shorter-duration segments of the curve during periods of market stress. Compared to the softmax allocation, the mean–variance strategy exhibits higher turnover, particularly during stressed market conditions, reflecting stronger sensitivity to changes in expected returns and covariance estimates.

6 Conclusion and Outlook

This thesis has examined whether systematic duration timing can deliver economically and statistically meaningful improvements over passive maturity allocation.

In the global WGBI setting, the evidence indicates that carry-based signals contain economically relevant information and can generate improvements relative to the equal-weighted benchmark. Among the factor-based approaches, the IC-based allocation rule emerges as the most defensible specification. By weighting signals according to their historical cross-sectional correlation with returns, the IC strategy combines economic interpretability with disciplined signal aggregation and delivers the most robust risk-adjusted performance within the factor framework. A pure carry strategy also represents a valid and simpler alternative, and in some specifications exhibits stronger statistical support. However, its more concentrated exposure makes it potentially less effective in protecting against drawdowns compared to the diversified IC-based approach.

In contrast, the U.S. Treasury market proves substantially more demanding. Once benchmark exposure and sampling uncertainty are properly accounted for, consistent statistical outperformance becomes difficult to achieve. In this setting, the equal-weighted allocation across maturity buckets remains a competitive and often preferable benchmark. While certain constrained machine-learning specifications and macro-based signals provide episodic improvements, none delivers strong and stable statistical evidence. For investors primarily concerned with capital preservation during stress regimes, threshold-based strategies can serve as protective overlays, but they are not reliable standalone alpha generators in benign environments.

The multi-country IPCA framework delivers the strongest overall results in the thesis. In particular, specifications combining categorical duration variables with a parsimonious softmax allocation rule produce economically meaningful gains and, under selected configurations, achieve statistical significance at conventional levels. These strategies function effectively as dynamic duration selectors, adjusting exposure across maturity segments in response to evolving conditions. Importantly, the IPCA-based allocations outperform the pure equal-weighted benchmarks considered in both the U.S.-only and WGBI settings, making this framework the most robust and consistent approach analyzed in the study.

The large number of models considered may raise concerns about potential data-snooping. However, the specifications are not the outcome of an unrestricted search over arbitrary models, but follow economically motivated choices regarding signals, allocation rules, and a limited set of parameter variations. In addition, statistical significance is not widespread across strategies, especially in the U.S. setting. This makes it unlikely that the results are driven by aggressive overfitting. Although model uncertainty cannot be fully ruled out, the structured and theory-driven design of the analysis helps contain the risk of spurious findings.

Several extensions naturally follow from this analysis. First, portfolio construction has been deliberately kept simple throughout the thesis in order to isolate the informational content of the signals. Exploring alternative portfolio construction schemes (such as risk-based allocations, volatility targeting, or optimization-based approaches) could improve the translation of signals into portfolio performance without altering the

underlying predictors.

Second, all strategies rebalance using discrete target weights, which can generate non-negligible turnover. Introducing weight-smoothing mechanisms or partial adjustment rules may help reduce transaction costs and improve net performance, particularly for strategies that react to rapidly changing signals.

Finally, the multi-country analysis based on IPCA could be extended beyond the linear framework considered here. Recent work interprets IPCA as a linear analogue of autoencoder models. Extending the current setting to non-linear architectures would allow for richer representations of time-varying risk exposures and may further improve duration-timing performance, albeit at the cost of increased model complexity and data requirements, which would require careful validation.

References

[1] Clifford S. Asness, Tobias J. Moskowitz, and Lasse Heje Pedersen. Value and momentum everywhere. *Journal of Finance*, 68(3):929–985, 2013. doi: 10.1111/jofi.12021.

[2] Guido Baltussen, Menno Martens, and Oliver Penninga. Factor investing in sovereign bond markets: Deep sample evidence. Ssrn working paper, Robeco Institutional Asset Management, 2021. URL: https://ssrn.com/abstract=3873863.

[3] Dirk G. Baur and Thomas K. McDermott. Is gold a safe haven? international evidence. *Journal of Banking & Finance*, 34(8):1886–1898, 2010. URL: https://doi.org/10.1016/j.jbankfin.2009.12.008.

[4] Diego Bianchi, Markus Buchner, and Alberto Tamoni. Bond risk premiums with machine learning. *Review of Financial Studies*, 34(2):1046–1089, 2021. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract$_i$d = 3232721.

[5] Jordan Brooks and Tobias J. Moskowitz. Yield curve premia. *SSRN Working Paper*, 2017. URL: https://ssrn.com/abstract=2956411.

[6] Antonio Caruso and Luca Coroneo. Does real-time macroeconomic information help to predict interest rates? *Journal of Money, Credit and Banking*, 55(8):2028–2058, 2023. URL: https://doi.org/10.1111/jmcb.13021.

[7] Anna Cieslak and Pavol Povala. Expected returns in treasury bonds. *Review of Financial Studies*, 28(10):2859–2901, 2015. URL: https://ssrn.com/abstract=1709636.

[8] John H. Cochrane and Monika Piazzesi. Bond risk premia. *American Economic Review*, 95(1):138–160, 2005. URL: https://doi.org/10.1257/0002828053828581.

[9] Francis X. Diebold and Canlin Li. Forecasting the term structure of government bond yields. *Journal of Econometrics*, 130(2):337–364, 2006. URL: https://doi.org/10.1016/j.jeconom.2005.03.005.

[10] Shihao Gu, Bryan T. Kelly, and Dacheng Xiu. Autoencoder asset pricing models. *Journal of Econometrics*, 222(1):429–450, 2021. URL: https://doi.org/10.1016/j.jeconom.2020.07.009.

[11] Brian Hurst, Yao Hua Ooi, and Lasse Heje Pedersen. A century of evidence on trend-following investing. *Journal of Portfolio Management*, 40(1):15–29, 2014. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2993026.

[12] Antti Ilmanen. Forecasting u.s. bond returns. *The Journal of Fixed Income*, 1997. URL: https://www.aqr.com/Insights/Research/Journal-Article/Forecasting-US-Bond-Returns.

[13] Bryan T. Kelly, Diogo Palhares, and Seth Pruitt. Modeling corporate bond returns. *Journal of Finance*, 78(4):2333–2388, 2023. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract$_i$d = 3720789.

[14] Bryan T. Kelly, Seth Pruitt, and Yinan Su. Instrumented principal component analysis. *Journal of Finance*, 74(6):2857–2914, 2019. URL: hhttps://papers.ssrn.com/sol3/papers.cfm?abstract$_i d = 2983919$.

[15] Bryan T. Kelly, Seth Pruitt, and Yinan Su. Characteristics are covariances: A unified model of risk and return. *Journal of Finance*, 76(6):2909–2955, 2021. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract$_i d = 3032013$.

[16] Marko Kolanovic and Zhenyu Wei. Momentum strategies across asset classes — risk factor approach to trend following. Technical report, J.P. Morgan, 2015. URL: https://www.cmegroup.com/education/files/jpm-momentum-strategies-2015-04-15-1681565.pdf.

[17] Oliver Ledoit and Michael Wolf. Robust performance hypothesis testing with the sharpe ratio. *Journal of Empirical Finance*, 15(5):850–859, 2008. URL: https://doi.org/10.1016/j.jempfin.2008.03.002.

[18] Andrew W. Lo. The statistics of sharpe ratios. *Financial Analysts Journal*, 58(4):36–52, 2002. URL: https://doi.org/10.2469/faj.v58.n4.2453.

[19] Sydney C. Ludvigson and Serena Ng. Macro factors in bond risk premia. *The Review of Financial Studies*, 22(12):5027–5067, 2009. URL: https://doi.org/10.1093/rfs/hhp081.

[20] Daniel L. Thornton and Giorgio Valente. Out-of-sample predictions of bond excess returns and forward rates: An asset allocation perspective. *The Review of Financial Studies*, 25(10):3141–3168, 2012. URL: https://academic.oup.com/rfs/article-abstract/25/10/3141/1573606.

Acknowledgments

This project has been conducted with the help of artificial intelligence tools, which have supported technical writing and code debugging/writing. However, all ideas, interpretations, mathematical derivations, methodological choices, and conclusions presented in this work remain the sole responsibility of the author.