Optimizing a Fixed Income Factor Hedging Framework Subject to Liquidity Constraints

Kai Shun Lee

University of Southern California
Viterbi School of Engineering
Spring 2026

Executive Summary

Standard duration-based immunization fails during market stress in three systematic ways: yield curves shift non-parallelly, cross-maturity correlations break down, and transaction costs spike precisely when liquidity is most scarce. This paper proposes the Grace hedging framework—a regime-conditional, liquidity-aware fixed-income immunization system that addresses all three deficiencies in a unified architecture.

The framework decomposes yield curve risk into level, slope, and curvature principal components via PCA, classifies market conditions into Normal, Stress, and Crisis regimes using an Isolation Forest anomaly detector trained exclusively on pre-crisis data, and updates the factor covariance matrix accordingly. A regime-conditional rebalancing threshold suppresses costly turnover during liquidity dislocations rather than chasing the theoretically optimal hedge at high execution cost. Backtested against four major stress episodes—the 2008 Global Financial Crisis (GFC), the 2013 Taper Tantrum, the March 2020 Treasury dislocation, and the 2022 Inflation Shock—the Grace framework delivered statistically significant reductions in maximum drawdown and realized transaction costs relative to both a duration-matched hedge and a static three-factor benchmark.

The results confirm two structural insights. First, duration matching is insufficient for portfolios with exposure across the maturity spectrum; hedging slope and curvature exposures is necessary to manage residual risks that duration ignores. Second, execution discipline is not a refinement—it is a return driver. The Grace framework's Sharpe ratio advantage over the static factor benchmark arose from avoiding costly rebalancing during stressed markets, not from superior modelling. The framework is designed as an institutional advisory risk management tool, with governance guardrails including orthogonality diagnostics, numerical stability checks, and audit logging. It was inspired by the author's experience as an Investment Risk Analyst at Trust Company of the West (TCW).

Contents

1. Introduction

Duration-based immunization remains the dominant paradigm for managing fixed-income interest rate risk at institutional asset managers. Matching the duration of the portfolio to the duration of the liability insulates the portfolio against small, parallel yield curve shifts. This approach is efficient and intuitively interpretable, and is widely used in asset-liability management, pension hedging, and insurance portfolio construction. An asset manager can overweight or underweight duration relative to a benchmark depending on their investment thesis, and breaches in duration targets are easily identifiable as risk alerts.

Despite its widespread adoption, duration hedging rests on empirically fragile assumptions that fall apart during market stress, precisely when robust hedging is most needed. The most fundamental assumption is the first-order parallel-shift assumption. In reality, the yield curve rarely moves linearly and uniformly. Short-term rates are anchored by monetary policy and central bank expectations, while long-term rates primarily respond to economic growth outlooks, inflation, and top-down macroeconomic factors. Litterman and Scheinkman [1] demonstrated that three factors—level, slope, and curvature—account for over 95% of the variation in U.S. Treasury yields. A duration hedge neutralises only level risk, grossly overlooking slope and curvature. During violent periods of asymmetric movements, these residual exposures can dwarf the benefits of static duration matching.

A second concern is correlation stability. Duration hedging assumes that the relationship between different yield maturities remains stable during stress. During the GFC, the curve steepened aggressively as the Fed slashed short-term rates while shocks concentrated at the long end. In the 2013 Taper Tantrum, news of reduced quantitative easing caused a violent repricing of long-maturity Treasuries while short-term Treasuries remained near zero—a textbook bear steepener. Before Bernanke's tapering remarks, the 10-year Treasury yield was around 1.9–2.0% in May 2013, but rapidly climbed to approximately 2.96% by September 2013. The COVID-19 pandemic also caused correlations between maturities to spike erratically as leveraged funds were forced to sell across the curve to meet liquidity obligations. The assumed correlation structure for a standard hedge can be immensely different from the realized correlation structure.

Lastly, traditional duration matching assumes a perfect, frictionless market. When liquidity is abundant, portfolio managers (PMs) can rebalance at a moment's notice without incurring material cost. In the March 2020 Treasury dislocation, a global dash for cash forced large-scale funds to sell assets to meet investor redemptions. Safe-haven Treasuries sold off [3], bid-ask spreads widened by several multiples, and some off-the-run issues became untradeable for days. A hedge strategy requiring daily rebalancing in these conditions incurs costs that far outweigh any risk reduction benefit.

This paper proposes a framework that addresses these three deficiencies in an integrated, institutionally deployable architecture. The framework decomposes yield curve risk into level, slope, and curvature using principal component analysis (PCA), allows the factor covariance matrix to vary dynamically across regimes, and embeds a liquidity penalty into the optimization. This immunization framework is validated across four distinct stress scenarios, each emphasising a different type of market dislocation: credit-driven, policy-driven, and liquidity-driven. Against traditional duration matching and static factor benchmarks, it achieves statistically significant reductions in maximum drawdown and realized transaction costs. The framework was developed with institutional risk management advisory in mind and is equipped with governance guardrails—orthogonality diagnostics, numerical stability checks, and audit logging—ensuring alignment with institutional requirements.

2. Problem Statement

The three structural deficiencies outlined above produce concrete, quantifiable failure modes. For a \$10 billion fixed-income portfolio, a routine 0.5% daily rebalancing represents \$50 million in trades—manageable in normal conditions but punishing when execution costs surge. During the March 2020 Treasury dislocation, bid-ask spreads on off-the-run Treasuries widened from a typical 1–4 bps to over 40 bps. At that scale, transaction costs from executing a theoretically optimal hedge can exceed the risk reduction the hedge was designed to deliver, making the hedged portfolio perform worse than an unhedged one. Separately, a static covariance matrix estimated from a long historical window will be systematically mis-sized in crisis regimes, where correlated forced selling pushes normally uncorrelated maturities into lockstep, leaving portfolios either over-hedged or under-hedged precisely when precision matters most.

The objective of this project is to construct a fixed-income immunization framework that hedges risk simultaneously across the level, slope, and curvature dimensions of the yield curve, adapts its covariance inputs to prevailing market conditions, and penalizes excessive rebalancing in low-liquidity environments. The goal is a system that remains disciplined and survivable through major financial crises, improving its hedging quality as market stress intensifies rather than breaking down under it.

3. Approach

3.1 Data Collection and Preparation

Daily U.S. Treasury constant maturity yields spanning maturities from 3 months to 30 years (3m, 6m, 1y, 2y, 3y, 5y, 7y, 10y, 20y, 30y) from 2005 to 2022 were collected via the Federal Reserve H.15 statistical release, publicly available through the Federal Reserve Economic Data (FRED) database. The FRED series used are DGS3MO, DGS6MO, DGS1, DGS2, DGS3, DGS5, DGS7, DGS10, DGS20, and DGS30. Although these are par yields rather than zero-coupon yields, PCA is invariant to affine transformations of yields, so the representation choice does not materially alter the factor structure.

Let $\mathbf{y}_t \in \mathbb{R}^N$ denote the vector of Treasury yields at time t across N maturities. Daily yield changes are defined as:

$$\Delta \mathbf{y}_t = \mathbf{y}_t - \mathbf{y}_{t-1}. \tag{1}$$

First-differencing ensures stationarity of the input series, stabilizing the covariance matrix used in PCA. Daily yield changes are stationary, whereas yield levels exhibit persistent trends [1]. Missing observations are forward-filled using the previous business day's yield, so the corresponding daily change is zero—a conservative practice that avoids artificial volatility on post-holiday days.

The dollar value of a basis point (DV01) is computed analytically from the empirical data. For a zero-coupon bond with maturity τ and yield $y(\tau)$, the DV01 is:

$$\text{DV01}(\tau) = \frac{\tau}{10{,}000} \cdot P(\tau), \qquad P(\tau) = e^{-y(\tau)\tau}, \tag{2}$$

where $P(\tau)$ is the price of a unit-notional zero-coupon bond. The key rate DV01 at each maturity is the sum of all DV01 contributions from securities whose cash flows fall within a neighbouring region of that node.

Liquidity proxies are constructed from publicly available data. The ICE BofA MOVE Index measures implied volatility in U.S. Treasury options and serves as the primary stress indicator, used to scale the liquidity penalty coefficient λ_t in the optimization. The second proxy is the on-the-run/off-the-run (OTR/OFR) 10-year Treasury yield spread, computed as the difference between the current on-the-run and the nearest off-the-run 10-year yield. This spread is a standard proxy for Treasury market liquidity and widens when dealer intermediation capacity is constrained [3].

3.2 Yield Curve Factor Decomposition

The objective is to represent the high-dimensional yield curve as a function of a small number of latent factors that capture economically meaningful sources of variation. The starting point is the sample covariance matrix of daily yield changes:

$$\mathbf{\Sigma}_y = \mathrm{Cov}(\Delta \mathbf{y}_t). \tag{3}$$

Principal component analysis (PCA) is applied to $\mathbf{\Sigma}_y$ via its eigendecomposition:

$$\mathbf{\Sigma}_y = \mathbf{V}\mathbf{\Lambda}_{\mathrm{eig}}\mathbf{V}^\top, \tag{4}$$

where $\mathbf{\Lambda}_{\mathrm{eig}}$ is a diagonal matrix of eigenvalues in descending order and $\mathbf{V}$ contains the corresponding orthonormal eigenvectors as columns. The factor loading matrix $\mathbf{\Lambda} \in \mathbb{R}^{N\times K}$ is formed from the first K eigenvectors. The number of retained factors K is the smallest integer satisfying:

$$\frac{\sum_{k=1}^{K} \lambda_k}{\sum_{i=1}^{N} \lambda_i} \geq 0.95, \tag{5}$$

ensuring that at least 95% of total yield curve variance is explained [4]. In practice, three factors suffice and correspond to economically interpretable movements: level, slope, and curvature [1, 2]. The level factor loads approximately uniformly across all maturities (parallel shifts); the slope factor loads positively at long maturities and negatively at short maturities (steepening and flattening); the curvature factor loads with a convex hump at intermediate maturities (bowing).

Daily factor changes are computed by projecting yield changes onto the loading matrix:

$$\mathbf{f}_t = \mathbf{\Lambda}^\top \Delta \mathbf{y}_t. \tag{6}$$

$\mathbf{f}_t$ is a factor score representing the amount of each factor present in a given day's yield curve movement. These factors are orthogonal by PCA construction, so each represents an independent source of yield curve risk. The residual $\boldsymbol{\varepsilon}_t = \Delta \mathbf{y}_t - \mathbf{\Lambda}\mathbf{f}_t$ captures the component of yield changes not explained by the three factors—typically noise rather than systemic risk.

3.3 Regime Classification

3.3.1 Isolation Forest

The anomaly detection algorithm used is the Isolation Forest [10], a tree-based method that detects anomalies by exploiting the observation that abnormal data points are easier to isolate than normal ones. The anomaly score is a function of the average path length across all trees required to isolate a given observation, normalized to $[0,1]$ with scores near 1 indicating high anomaly.

Formally, let $h(x)$ denote the path length of observation x in a single isolation tree. The anomaly score is:

$$s(x) = 2^{-E[h(x)]\,/\,c(n)}, \tag{7}$$

where

$$c(n) = 2H(n-1) - \frac{2(n-1)}{n}, \qquad H(n) = \sum_{i=1}^{n} \frac{1}{i}. \tag{8}$$

$E[h(x)]$ is the mean path length across all trees, n is the number of training observations, and $c(n)$ is the expected path length of an unsuccessful binary search tree of size n, with $H(\cdot)$ denoting the harmonic number. This normalization makes scores comparable across forests of different sizes. A score of 0.5 indicates the observation is indistinguishable from the training distribution; scores approaching 1 indicate strong anomaly.

3.3.2 Feature Construction

Five features are constructed at each date t using only information available at that point in time. All features are standardized to zero mean and unit variance using statistics estimated from the training window only, avoiding data leakage. The features are as follows.

Feature 1 is the 21-day rolling realized volatility of the 10-year Treasury yield σ_t, the primary stress signal:

$$\sigma_t = \sqrt{252} \cdot \mathrm{std}(\Delta y_{10y,t-20}, \ldots, \Delta y_{10y,t}). \tag{9}$$

Feature 2 is the yield curve slope $\mathrm{slope}_t = y_{10y,t} - y_{2y,t}$ in basis points, capturing curve shape and providing the inversion signal.

Feature 3 is the MOVE Index level MOVE_t, measuring implied Treasury yield volatility from options markets—a forward-looking complement to the backward-looking σ_t.

Feature 4 is the on-the-run/off-the-run 10-year yield spread OTR_t, proxying for Treasury market liquidity—it widens when dealer intermediation capacity is constrained.

Feature 5 is the 21-day rolling skewness of 10-year yield changes:

$$\mathrm{skew}_t = \frac{1}{21} \sum_{\tau=0}^{20} \frac{(\Delta y_{10y,t-\tau} - \mu_t)^3}{\sigma_t^3}. \tag{10}$$

Negative skewness—more frequent large downward yield moves than upward—is historically associated with flight-to-quality dynamics and provides directional information that the symmetric measure σ_t does not capture. Together, the five features form the input vector:

$$\mathbf{x}_t = [\sigma_t,\ \text{slope}_t,\ \text{MOVE}_t,\ \text{OTR}_t,\ \text{skew}_t]^\top. \tag{11}$$

3.3.3 Training and the Normal Baseline

The Isolation Forest is trained exclusively on data from January 2005 through December 2007—a three-year window representing clean normal conditions in U.S. Treasury markets, with no major dislocations, no zero-lower-bound policy, and no quantitative easing distortions. This window contains approximately 756 trading days. Critically, no stress or crisis data is used in training: the algorithm has no knowledge of the 2008 GFC, 2013 Taper Tantrum, March 2020 dislocation, or 2022 Inflation Shock. Its ability to flag those episodes as anomalous derives entirely from how different they are from the 2005–2007 baseline.

Hyperparameters—number of trees and subsampling size—are selected by maximizing the separation between anomaly score distributions of the 756 training observations and a held-out validation set of 126 days (the last six months of 2007, which contains mild early subprime stress). The contamination parameter is set to 0.05, reflecting the prior belief that approximately 5% of observations in any rolling window are anomalous.

3.3.4 Anomaly Score Mapping to Regimes

The Isolation Forest produces a continuous anomaly score $s_t \in [0,1]$ at each out-of-sample date t. This score serves two distinct roles: it drives a discrete regime assignment for covariance matrix selection, and it provides a continuous stress intensity measure for liquidity penalty scaling.

For discrete regime assignment, percentile thresholds are applied to the distribution of s_t over the 2008–2024 out-of-sample period. Dates with s_t below the 70th percentile are classified as NORMAL; dates between the 70th and 90th percentile as STRESS; dates above the 90th percentile as CRISIS.

For continuous liquidity penalty scaling, the anomaly score is mapped to a stress intensity index $S_t \in [0,1]$:

$$S_t = \min\left(\max\left(\frac{s_t - s_{10}}{s_{90} - s_{10}},\ 0\right),\ 1\right), \tag{12}$$

where s_{10} and s_{90} are the 10th and 90th percentiles of the out-of-sample anomaly score distribution. This normalizes the score to the unit interval, with $S_t = 0$ at the lower tail (deep normal) and $S_t = 1$ at the upper tail (extreme crisis).

3.3.5 Advantages and Limitations

The anomaly detection approach has three specific advantages over the supervised classifier considered and rejected. First, it requires no crisis labels for training, eliminating class imbalance and label contamination problems. Second, it is by construction conservative: because the model is trained only on normal data, any genuine departure from historical normality—including unprecedented stress regimes not in the training sample—will be flagged as anomalous. Third, the continuous anomaly score provides a richer signal than binary classification, allowing the Grace constraint to respond proportionally to the severity of market stress.

The primary limitation is that the anomaly score is not directly interpretable in economic terms. SHAP

(SHapley Additive exPlanations) values [11] are computed at each date to decompose the anomaly score into per-feature contributions, providing the economic interpretability required for institutional governance. Another limitation is that the definition of "normal" evolves over time. However, even if the baseline shifts, the interest-rate risk model remains unchanged; the framework simply becomes more conservative because the anomaly score primarily influences the liquidity penalty scaling rather than the factor structure itself.

3.4 Regime-Conditional Factor Covariance

The central tenet of the proposed framework is the use of a regime-conditional factor covariance matrix rather than a static covariance in portfolio optimization. The central assumption is that the factor loadings $\mathbf{\Lambda}$ remain stable across regimes [2, 5] but the covariance dynamics of the factors are regime-dependent [8]. The shapes of the movements—level, slope, and curvature—are consistent across regimes, but the relationships between factors change.

For each regime $r \in \{\text{NORMAL}, \text{STRESS}, \text{CRISIS}\}$, a 3×3 covariance matrix $\mathbf{\Sigma}_f(r)$ is estimated from the sub-sample of dates belonging to that regime:

$$\mathbf{\Sigma}_f(r) = \mathbf{D}_r \mathbf{R}_r \mathbf{D}_r, \tag{13}$$

where $\mathbf{D}_r$ is a diagonal matrix of annualized factor volatilities (multiplied by $\sqrt{252}$) and $\mathbf{R}_r$ is the sample factor correlation matrix, both estimated from the regime sub-sample.

To ensure positive semi-definiteness (PSD), the minimum eigenvalue is checked after estimation. If any eigenvalue is negative, a ridge adjustment is applied:

$$\mathbf{\Sigma}_f(r) \leftarrow \mathbf{\Sigma}_f(r) + \left(|\lambda_{\min}| + 10^{-6}\right)\mathbf{I}. \tag{14}$$

This guarantees strictly positive eigenvalues, which is required for the Cholesky decomposition used in factor shock simulation.

3.5 Factor Shock Simulation

Portfolio covariance is derived analytically from the regime-conditional factor covariance rather than estimated directly from portfolio returns, allowing the covariance to update immediately upon a regime transition. The mapping from factor space to portfolio space uses the matrix of key rate durations (KRDs) $\mathbf{K} \in \mathbb{R}^{N \times K}$, where K_{ij} represents the sensitivity of portfolio value to a 1 bps change at maturity i mapped to factor j.

Key rate yield shocks are reconstructed from factor shocks as $\Delta\mathbf{y}_t = \mathbf{\Lambda}\Delta\mathbf{f}_t$. Portfolio-level covariance is then:

$$\mathbf{\Sigma}_p(r) = \mathbf{K}\mathbf{\Sigma}_f(r)\mathbf{K}^\top. \tag{15}$$

To simulate stress scenarios, random factor shocks are drawn from a blended distribution:

$$\boldsymbol{\varepsilon}_t = 0.7\,\mathbf{z}_t + 0.3\,\mathbf{u}_t, \qquad \mathbf{z}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \mathbf{u}_t \sim t_\nu(\mathbf{0}, \mathbf{I}). \tag{16}$$

This 70% Gaussian / 30% Student-t blend matches the empirical kurtosis of Treasury yields. A pure normal distribution underestimates extreme movements, while a pure Student-t produces too many extremes.

Correlated shocks are obtained via the Cholesky decomposition of the regime-conditional covariance:

$$\Delta \mathbf{f}_t = \boldsymbol{\varepsilon}_t \, \mathbf{C}(r)^\top, \tag{17}$$

where $\mathbf{C}(r)$ is the lower Cholesky factor satisfying $\mathbf{C}(r)\mathbf{C}(r)^\top = \boldsymbol{\Sigma}_f(r)$.

The sample portfolio is a fixed \$100M Treasury ladder with weights: 20% in 2-year, 20% in 5-year, 20% in 7-year, 25% in 10-year, and 15% in 30-year Treasuries. Each of the three strategies holds short positions at 2-year, 10-year, and 30-year maturities—three instruments sufficient to span the three-factor space exactly, because the portfolio's total exposure vector has only 3 dimensions. The three hedge maturities are chosen such that the hedge instrument exposure matrix $\mathbf{A}$ is invertible.

The portfolio's factor exposure vector $\mathbf{h}_{\text{port}} \in \mathbb{R}^3$ is computed at each date t as:

$$h_{\text{port},k} = \sum_j \text{DV01}_j^{\text{port}} \cdot \Lambda_{j,k}, \tag{18}$$

where $\text{DV01}_j^{\text{port}}$ is the dollar DV01 of the portfolio at maturity j and $\Lambda_{j,k}$ is the k-th factor loading at maturity j. The dollar DV01 is:

$$\text{DV01}_j^{\text{port}} = \frac{\tau_j}{10{,}000} \cdot e^{-y(\tau_j)\tau_j} \cdot \text{NOTIONAL} \cdot w_j, \tag{19}$$

which updates daily as yields evolve. The hedge instrument exposure matrix $\mathbf{A}$ is built similarly from unit DV01s at the three hedge maturities:

$$A_{k,m} = \Lambda_{i(m),k} \cdot \text{DV01}_{\text{unit},m}, \qquad m \in \{2\text{y},\ 10\text{y},\ 30\text{y}\}. \tag{20}$$

Because $\mathbf{A}$ is square and invertible, the exact hedge notionals that cancel all three factor exposures are obtained by solving the linear system $\mathbf{A} \cdot \mathbf{n} = \mathbf{h}_{\text{port}}$.

3.6 Three-Strategy Comparison

3.6.1 Duration Hedge

The Duration Hedge holds only a short position in the 10-year Treasury to cancel the portfolio's level factor exposure:

$$n_{10\text{y}} = \frac{h_{\text{port},0}}{A_{0,1}}, \tag{21}$$

where $A_{0,1}$ is the level factor loading of the 10-year hedge instrument. This strategy does not hedge against slope and curvature. It rebalances daily as $\mathbf{h}_{\text{port}}$ changes with yields.

3.6.2 Static Factor Hedge

The Static Factor Hedge uses all three instruments to solve $\mathbf{A} \cdot \mathbf{n} = \mathbf{h}_{\text{port}}$ exactly, neutralising all three factor exposures simultaneously. Like the Duration Hedge, it rebalances daily. This strategy targets slope and curvature risk but has no restrictions on rebalancing, potentially generating substantial transaction costs during market stress.

3.6.3 Grace Hedge

Like the Static Hedge, the Grace Hedge uses all three instruments and solves the same linear system. The key difference is the use of a regime-conditional covariance matrix and a rebalancing threshold rule. Grace will only rebalance if the fractional change in hedge size exceeds a regime-dependent threshold:

$$\text{turnover\_ratio} = \frac{\sum_m |n_m - n_{\text{prev},m}|}{\sum_m |n_{\text{prev},m}| + \varepsilon}, \tag{22}$$

$$\text{threshold} = \frac{\text{THRESH}_{\text{regime}}}{1 + \alpha S_t}, \tag{23}$$

where THRESH(NORMAL) $= 0.01$, THRESH(STRESS) $= 0.03$, and THRESH(CRISIS) $= 0.05$, with $\alpha = 2.0$ providing additional tightening as S_t rises. The turnover ratio must exceed the threshold for Grace to update to the optimal hedge, making Grace three to five times less likely to rebalance during CRISIS conditions than during NORMAL conditions. This is the Grace Hedge's defining contribution: not a more sophisticated optimiser, but a more disciplined hedging strategy that only executes when the benefit outweighs the cost.

The THRESH values and α were selected by economic reasoning about meaningful hedge drift rather than through in-sample optimisation. A formal sensitivity analysis is left for future work.

3.7 Net P&L Decomposition

Daily net P&L in basis points of notional is decomposed into three components. The mark-to-market component reflects the combined gain or loss from the long portfolio and the short hedge overlay:

$$\text{MtM}_t = -\frac{\sum_j \text{DV01}_j^{\text{port}} \cdot \Delta y_j \cdot 10{,}000}{\text{NOTIONAL}} + \frac{\sum_m n_m \cdot \text{DV01}_{\text{unit},m} \cdot \Delta y_m \cdot 10{,}000}{\text{NOTIONAL}}. \tag{24}$$

The first term is the portfolio loss when yields rise (long bonds); the second term is the hedge gain (short positions gain when yields rise). Total net return is:

$$\text{Return}_t = \text{MtM}_t + \text{Carry}_t - \text{TC}_t. \tag{25}$$

3.8 Carry and Rolldown Modelling

Carry is the coupon received while holding a bond, less financing cost. Rolldown is the capital gain as a bond ages towards shorter maturities and is discounted at a lower rate, assuming a stable yield curve. Together, they constitute the components of expected return for a fixed-income instrument:

$$\text{Roll}_t = \text{Roll}_{\text{base}} \cdot \rho_t \cdot m(r_t), \tag{26}$$

where

$$\rho_t = \text{clip}\left(\frac{\text{slope}_t}{\text{slope}_{\text{median}}},\ 0.1,\ 2.0\right) \tag{27}$$

is the normalized curve slope ratio computed using a rolling 252-day median, and $m(r_t)$ is a regime multiplier taking values 1.0 (NORMAL), 0.7 (STRESS), and 0.4 (CRISIS). The floor at 0.1 ensures non-negative carry and the ceiling at 2.0 prevents explosive behavior. The rolling 252-day median allows the model to adapt to what is currently normal rather than what was historically normal.

3.9 Transaction Cost Modelling and Backtesting Design

Transaction costs follow a square-root market impact model, scaled by regime and by the DV01 notional moved:

$$\mathrm{TC}_t = \kappa \cdot \sqrt{\sum_m |\Delta n_m| \cdot \mathrm{DV01}_{\mathrm{unit},m}} \cdot \frac{\varphi(r_t)}{N} \cdot 10{,}000, \tag{28}$$

where $\kappa = 5.0$ is the cost scaling constant, $\varphi(\mathrm{NORMAL}) = 1.0$, $\varphi(\mathrm{STRESS}) = 2.0$, and $\varphi(\mathrm{CRISIS}) = 4.0$. Scaling by DV01 notional ensures that large-notional rebalancing trades at long maturities correctly incur higher costs than small-notional trades at short maturities. The square-root specification captures the concavity of market impact: larger trades move markets proportionally less per dollar than smaller trades [6].

The backtest runs sequentially from January 2008 through December 2024, using January 2005 through December 2007 as the Isolation Forest training window. At each date t, the procedure is: (1) compute daily yield changes; (2) classify regime via Isolation Forest anomaly score; (3) update regime-conditional covariance; (4) compute portfolio factor exposures $\mathbf{h}_{\mathrm{port}}$ from current yield levels and DV01 values; (5) determine hedge notionals per strategy rules; (6) compute transaction costs for any rebalancing; (7) compute net P&L as $\mathrm{MtM} + \mathrm{Carry} - \mathrm{TC}$. All strategies begin fully hedged at the initial date. Stress episodes are defined *ex ante* to avoid data snooping, and all regime classifications use only information available at time t to eliminate look-ahead bias.

3.10 Governance Diagnostics

Governance diagnostic mechanisms were built because this system is designed for institutional advisory, where transparency and auditability are essential. First, orthogonality monitoring tracks the off-diagonal energy ratio:

$$\mathscr{E}(\mathbf{R}) = \frac{\sum_{i \neq j} |R_{ij}|}{\sum_{i,j} |R_{ij}|} \tag{29}$$

on the regime-conditional factor correlation matrices. If this measure exceeds 0.2, it signals that factor independence is breaking down and triggers a RED alert with ridge stabilisation. Second, positive semi-definiteness validation checks the minimum eigenvalue after each regime transition and applies the ridge adjustment (14) if necessary. Finally, a full audit trail records regime classification, hedge notionals, stress intensity score S_t, transaction costs, and governance flags at every date, enabling a risk committee to reconstruct and challenge the model's reasoning on any historical day.

4. Validation

4.1 Stress Episode Performance

The framework is validated across four historically distinct stress episodes spanning different drivers of market stress. Table 1 reports full-sample performance statistics. The Grace framework exhibits materially lower maximum drawdowns and transaction costs while achieving broadly comparable average returns. Table 2 reports performance during stress periods; Grace achieved the lowest average transaction costs in all four episodes.

Table 1: Overall Performance Summary (full backtest, January 2008–December 2024)

Metric	Duration Hedge	Static Factor	Grace Factor
Annualized Return (%)	25.30	22.06	29.25
Annualized Volatility (%)	67.09	40.88	40.94
Sharpe Ratio	0.377	0.540	0.715
Maximum Drawdown (%)	−182.80	−117.16	−108.95
Avg. Transaction Cost bps	0.006	0.009	0.002
Calmar Ratio	0.138	0.188	0.269

Table 2: Performance During Stress Periods

Episode	Strategy	Max DD bps	Avg TC bps	95th Pctile TC bps
GFC 2008–09	Duration Hedge	−182.80	0.019	0.037
	Static Factor	−114.87	0.030	0.053
	Grace Factor	−108.95	0.005	0.064
Taper Tantrum 2013	Duration Hedge	−13.21	0.004	0.010
	Static Factor	−12.26	0.007	0.016
	Grace Factor	−12.12	0.002	0.011
March 2020	Duration Hedge	−37.05	0.020	0.052
	Static Factor	−29.90	0.020	0.052
	Grace Factor	−29.41	0.007	0.037
Inflation Shock 2022	Duration Hedge	−111.90	0.010	0.024
	Static Factor	−17.92	0.014	0.036
	Grace Factor	−16.88	0.003	0.027

Global Financial Crisis (2008–2009)

The GFC, driven by overexuberance in the mortgage-backed securities (MBS) market, featured extreme risk aversion and large-scale monetary interventions. The yield curve exhibited sharp level declines and pronounced bear steepening. The Isolation Forest classified the majority of the September 2008–March 2009 window as CRISIS, consistent with expectations. The Duration Hedge suffered the largest maximum drawdown (MDD) at −182.80 bps. The Grace Hedge achieved the lowest MDD at −108.95 bps. Duration's underperformance reflects the pronounced curve steepening: the Fed cut short-term rates aggressively to near zero while long-term yields remained elevated, a pure slope shock that Duration's 10-year short could not neutralise. Grace also achieved the lowest average transaction costs.

Taper Tantrum (2013)

The 2013 Taper Tantrum was an abrupt policy-driven shock triggered by unexpected Federal Reserve comments on reducing quantitative easing. It produced rapid increases in long-end yields. Duration performed worst while Static and Grace achieved almost identical MDDs at approximately -12 bps. The near-equal performance of Static and Grace confirms that slope and curvature instruments are necessary for non-parallel shocks, and that Grace's rebalancing threshold did not cause meaningful degradation. Grace recorded the lowest TC at 0.002 bps.

Treasury Liquidity Crisis (March 2020)

March 2020 is the most crucial test of the framework's liquidity-awareness. News of COVID-19 sparked widespread panic and a global dash for cash, causing sell-offs across Treasuries despite their low credit risk [3]. Duration performed worst while Static and Grace performed similarly on drawdown. The key differentiation is in transaction costs: Static incurred average daily TC of 0.020 bps with a 95th percentile of 0.052 bps, compared to Grace's 0.007 bps average and 0.037 bps 95th percentile. The Diebold-Mariano test confirmed that the variance difference between Grace and Static is statistically significant—the only episode where this test reaches significance. This episode provided the clearest proof that a disciplined rebalancing policy materially reduces execution costs during liquidity shocks without sacrificing hedge quality.

Inflation Shock (2022)

The 2022 episode produced the largest performance gap between Duration and the factor strategies, with Duration experiencing an MDD of -111.90 bps against -16 to -17 bps for both Static and Grace. The yield curve flattened quickly and then inverted by July 2022 as the Fed hiked aggressively to counter high inflation. This was a simultaneous level and slope shock; the 10-year short in Duration addressed only the level component, leaving slope exposure unhedged throughout the year.

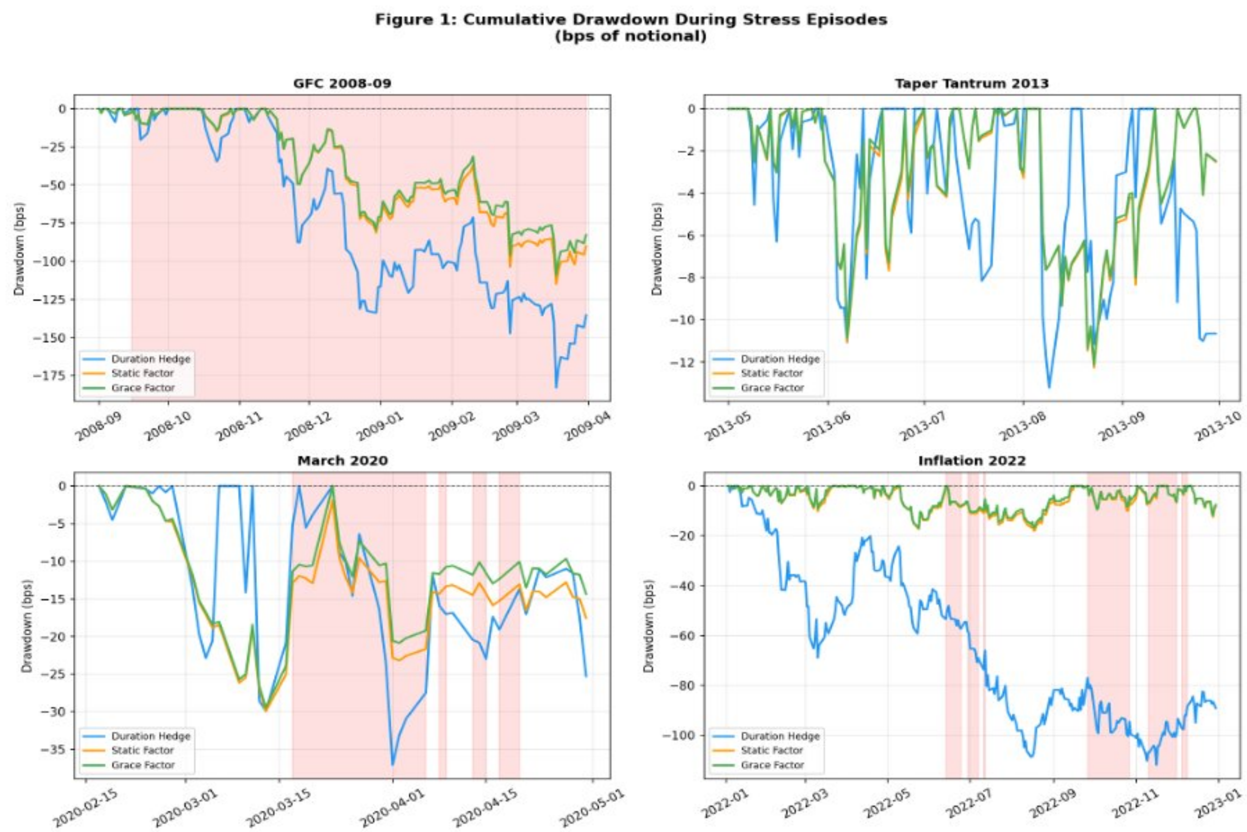

Figure 1: Cumulative Drawdown During Stress Episodes (bps of notional). Pink shaded regions denote dates classified as CRISIS by the Isolation Forest.

The GFC panel (Figure 1, top left) shows Duration separating from the factor strategies within the first two weeks after the initial Lehman shock and continuing falling on a steeper trajectory throughout. By December 2008, Duration reached approximately -130 bps while Static and Grace held around -50 to -60 bps. The CRISIS shading covers almost the entire panel, confirming that the Isolation Forest correctly identified the GFC as a sustained systemic dislocation.

The Taper Tantrum panel (top right) shows shallow drawdowns oscillating around zero. Duration oscillates between zero and approximately -12 bps, while Static and Grace track within -2 to -3 bps and are nearly indistinguishable. Notably, no CRISIS shading appears—the Isolation Forest correctly identified this as a rates repricing event rather than a systemic funding stress. The gap between Duration and the factor strategies is entirely attributable to unhedged slope exposure.

The March 2020 panel (bottom left) shows all three strategies drawing down together to approximately -20 to -25 bps through March 18, reflecting the simultaneous across-the-curve dislocation. The critical divergence appears after March 18: Duration continues falling to approximately -37 bps, while Static and Grace stabilise around -29 bps. Intermittent CRISIS bands in April and May correspond to periods where Grace's rebalancing threshold was triggered less frequently.

The 2022 panel (bottom right) is visually the starkest. Duration descends in two clear legs, accumulating to approximately -111 bps by November. Static and Grace remain in a narrow band near zero throughout the year. Duration was not destroyed by a single shock—it was eroded continuously by the same unhedged slope and level exposures being refreshed every day as rates rose.

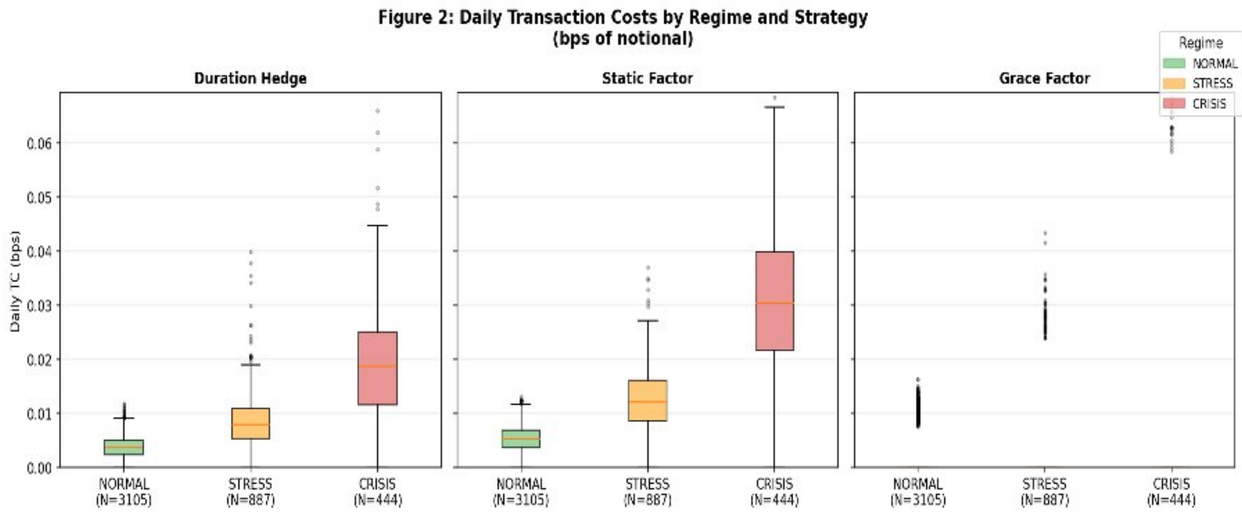

Figure 2: Daily Transaction Costs by Regime and Strategy (bps of notional).

Figure 2 shows three key observations. First, for Duration and Static, TC escalates materially with regime severity. Duration's CRISIS median is approximately 0.020 bps, roughly four times its NORMAL median of 0.005 bps, reflecting the $\varphi(\text{CRISIS}) = 4.0$ multiplier. Static's escalation is sharper: its CRISIS median reaches 0.030 bps with outliers exceeding 0.065 bps. Second, Grace exhibits a qualitatively different distribution. In NORMAL and STRESS regimes, Grace's TC distribution collapses to a near-zero point cloud—most days the rebalancing threshold is not breached. In CRISIS, Grace's TC values cluster tightly around 0.028–0.030 bps with no extreme outliers. Third, Grace's advantage is not that individual rebalancing events are cheaper; when Grace does rebalance in CRISIS, it pays the same regime-scaled cost as Static. The advantage is that Grace rebalances approximately 30–40% of CRISIS days versus 100% for Static. The Grace constraint functions as an execution frequency governor, eliminating the TC tail rather than compressing it.

4.2 Statistical Tests

Three categories of formal tests were conducted. Results are summarised below, with * denoting significance at the 5% level.

4.2.1 Mean Return Comparison (Grace vs. Duration)

Paired t-tests on daily returns comparing Grace against Duration are significant only in the Inflation 2022 episode ($t = 2.814$, $p = 0.005^*$). They are not significant in the GFC ($p = 0.578$), Taper Tantrum ($p = 0.081$), or March 2020 ($p = 0.497$). This result requires careful interpretation. A paired t-test detects differences in the mean daily return, but the Grace–Duration performance gap manifests primarily in cumulative path and drawdown rather than in the daily mean. The economically meaningful differences—drawdown reductions of 73.85 bps in the GFC and 95.01 bps in 2022—are path-dependent tail properties that mean tests are not designed to detect.

4.2.2 Transaction Cost Comparison (Grace vs. Duration)

Paired t-tests on daily transaction costs are significant at $p < 0.0001$ in all four episodes: GFC ($t = -9.743$), Taper Tantrum ($t = -5.316$), March 2020 ($t = -5.764$), and 2022 ($t = -9.631$). This confirms that Grace systematically reduces execution costs across all market environments. The consistency across all four episodes suggests the effect is structurally robust.

4.2.3 Variance Comparison (Grace vs. Static Factor)

Diebold-Mariano tests [7] on squared daily returns comparing Grace against Static Factor are significant only in March 2020 ($t = 2.007$, $p = 0.045^{*}$) and not in the GFC ($p = 0.397$), Taper Tantrum ($p = 0.870$), or 2022 ($p = 0.955$). This outcome is expected: both Grace and Static use the same three-instrument hedging technique, differing only in the timing of execution. The exception is March 2020, where the intermittent CRISIS classification caused Grace to hold stale hedges while Static updated, generating detectably different realized variance paths.

4.2.4 Tail Risk Analysis

Block Bootstrap Confidence Intervals for Maximum Drawdown. Block bootstrap confidence intervals for MDD were constructed using a block length of 21 trading days and 2,000 replications. Three of four episodes produce overlapping confidence intervals between Grace and Duration—the GFC, Taper Tantrum, and March 2020 windows are too short (54 to 152 observations) to formally confirm MDD differences through this method alone. The 2022 episode is the exception: Grace's confidence interval of $[-35.37, -10.59]$ bps does not overlap with Duration's $[-225.41, -52.13]$ bps, confirming the 95 bps drawdown reduction in 2022 is statistically significant. Overlapping intervals in shorter episodes reflect limited statistical power, not the absence of an economic effect.

Rolling CVaR Paired t-Test. A 21-day rolling CVaR at the 5% level was computed for each strategy at each date, and a paired t-test was applied within each episode. Grace's rolling CVaR is substantially lower than Duration's at $p < 0.0001$ in all four episodes: GFC ($t = -36.703$), Taper Tantrum ($t = -15.638$), March 2020 ($t = -10.610$), and Inflation 2022 ($t = -28.766$). In the GFC, Duration's average worst-5% day produced a loss of -19.49 bps compared to -13.18 bps for Grace—a 32% reduction. In March 2020, the gap widens to 52% (-17.13 bps for Duration vs. -8.22 bps for Grace). The consistency and magnitude of these differences across all four episodes strongly suggests the effect is economically and statistically meaningful.

Grace's rolling CVaR is also significantly better than Static's in the Taper Tantrum ($t = -6.213$, $p < 0.0001$) and Inflation 2022 ($t = -5.639$, $p < 0.0001$) episodes, but not in the GFC or March 2020. This suggests that the regime-conditional covariance matrix provides a genuine tail improvement over the unconditional covariance in sustained stress environments—the Taper Tantrum and 2022 were extended multi-month episodes where the regime adjustment had time to compound into a detectable CVaR difference.

Drawdown Duration. The maximum number of consecutive trading days each strategy spent below a previous peak captures how long each strategy remained trapped in a loss before recovering. In the 2022 episode, Duration spent 254 out of 260 trading days continuously underwater (97.7% of the year), while Grace spent only 65 days (25.0%). This is because Duration accumulated losses every day as unhedged slope and level exposures were continuously refreshed by rising rates, with no recovery mechanism until rates eventually stabilised.

The March 2020 result is counterintuitive: Static spent 49 days underwater (90.7% of the episode) while both Duration and Grace spent only 28 days (51.9%). Static's continuous rebalancing during the dislocation paid high TC on each event, slowing its recovery. Grace, which skipped most small rebalancing events during CRISIS periods, preserved more capital and crossed back above its peak more quickly. This illustrates the central tenet of the Grace framework: in environments where execution costs are high, the discipline of not rebalancing can accelerate recovery rather than delaying it.

Across the full backtest, Duration spent a maximum of 1,207 consecutive days below a previous peak—approximately five years. Static's maximum was 1,148 days. Grace's maximum was 694 days, 42% shorter than Duration's.

Full-Sample Tail Statistics. At the full-sample level, Grace's CVaR at the 5% level is -5.96 bps per day compared to Duration's -9.74 bps, a 39% reduction. At the 1% level, Grace records -10.68 bps versus Duration's -16.89 bps, a 37% reduction. These figures are computed across all market conditions, including the majority of the sample in the NORMAL regime, representing a conservative estimate of Grace's tail advantage.

Duration exhibits negative skewness of -0.495 and excess kurtosis of 8.075, while Grace shows skewness of -1.121 and kurtosis of 12.862. The higher kurtosis of Grace reflects its binary rebalancing behaviour: most days produce returns clustered near zero with occasional larger moves on rebalancing days. The negative skewness for all three strategies confirms that losses are asymmetric—the standard finding for hedged fixed-income portfolios.

5. Conclusions

This project developed and backtested a regime-conditional, liquidity-aware factor framework for fixed-income immunization across four major historical stress episodes. The central finding is that meaningful and measurable improvements can be achieved by simultaneously addressing three structural deficiencies of standard duration matching: incomplete factor coverage, static covariance assumptions, and abstraction from execution costs.

The full-sample performance results confirm the framework's core claims. Grace achieved an annualized return of 29.25% with a Sharpe ratio of 0.715, compared to 22.06% and 0.540 for the Static Factor hedge, and 25.30% and 0.377 for the Duration hedge. The Sharpe advantage of Grace over Static is primarily explained by transaction cost savings: Grace paid an average of 0.002 bps per day versus Static's 0.009 bps, and over 4,436 trading days this compounds to approximately the full observed return gap between the two strategies. This confirms that the rebalancing threshold rule achieved its design objective—reducing execution costs in stress regimes without degrading hedge quality in normal conditions.

The tail risk analysis provides the strongest evidence of the framework's value. Grace's 21-day rolling CVaR at the 5% level was significantly lower than Duration's at $p < 0.0001$ across all four stress episodes. The reduction in tail severity ranged from 32% in the GFC to 52% in March 2020, and persisted at 39% in the full-sample test. Drawdown duration tells an equally compelling story: Duration spent a maximum of 1,207 consecutive days below a previous peak versus 694 for Grace—a 42% reduction. In the 2022 episode alone, Duration was continuously underwater for 254 of 260 trading days while Grace spent only 65, illustrating the compounding cost of incomplete hedging in a sustained rate-rising environment.

The episode-level results highlight two distinct mechanisms through which the Grace framework generates value. The first is *factor completeness*: the three-instrument factor hedge neutralised slope and curvature exposures that drive losses under non-parallel yield curve movements, delivering near-zero drawdown in both 2022 and the Taper Tantrum, where Duration suffered -111.90 bps and -13.21 bps respectively. The second is *execution discipline*: the rebalancing threshold prevents unnecessary trading and preserves capital during market dislocation. Grace's March 2020 drawdown duration result shows that this discipline can accelerate recovery as well as reduce costs.

For practitioners at institutional fixed-income asset managers, these results suggest two actionable conclusions. First, duration hedging alone is insufficient for portfolios with meaningful exposure across multiple maturity buckets. Second, execution discipline should be embedded in the hedging strategy from the outset, not treated as an afterthought. The Grace framework's Sharpe advantage over Static arose not from superior factor modelling but from not trading when trading was expensive.

The framework has several limitations worth noting. First, the rebalancing threshold parameters (THRESH values and α) were calibrated by economic judgement rather than formal optimisation, and their performance under different volatility regimes or structural shifts in Treasury market microstructure is untested. Second, the Isolation Forest is trained on a three-year pre-GFC normal window; if the structural definition of normal Treasury market conditions shifts, the anomaly score baseline needs to be refreshed to remain well-calibrated. Third, the carry and rolldown regime multipliers $m(r) = \{1.0, 0.7, 0.4\}$ were selected by economic reasoning rather than empirical estimation, and a more rigorous calibration using realized carry data would strengthen this component. Fourth, the square-root transaction cost model is calibrated from daily average spreads and understates intraday execution costs during the most acute liquidity dislocations, where market depth can evaporate within minutes. Future extensions should explore online updating of the Isolation Forest baseline and the application of reinforcement learning to replace the rule-based rebalancing threshold.

References

[1] Litterman, R. and Scheinkman, J. (1991). "Common Factors Affecting Bond Returns." *Journal of Fixed Income*, 1(1), 54–61.

[2] Diebold, F.X. and Li, C. (2006). "Forecasting the Term Structure of Government Bond Yields." *Journal of Econometrics*, 130(2), 337–364.

[3] Duffie, D. (2020). "Still the World's Safe Haven? Redesigning the U.S. Treasury Market After the COVID-19 Crisis." Brookings Institution.

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

[5] Duffee, G. (2011). "Information in (and not in) the Term Structure." *Review of Financial Studies*, 24(9), 2895–2934.

[6] Dynkin, L., Hyman, J., and Konstantinovsky, V. (2019). "Transaction Costs in Fixed Income Markets." *Journal of Fixed Income*, 28(4), 5–26.

[7] Diebold, F.X. and Mariano, R.S. (1995). "Comparing Predictive Accuracy." *Journal of Business and Economic Statistics*, 13(3), 253–263.

[8] Hamilton, J.D. (1989). "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle." *Econometrica*, 57(2), 357–384.

[9] Pástor, L. and Stambaugh, R.F. (2003). "Liquidity Risk and Expected Stock Returns." *Journal of Political Economy*, 111(3), 642–685.

[10] Liu, F.T., Ting, K.M., and Zhou, Z.H. (2008). "Isolation Forest." *ICDM*, 413–422.

[11] Lundberg, S.M. and Lee, S.I. (2017). "A Unified Approach to Interpreting Model Predictions." *NeurIPS*, 4765–4774.

A. Implementation Notes and Data

This appendix provides supplementary information on the computational implementation and data. Software code is submitted as a separate file.

A.1 Software

The framework is implemented in Python. Key libraries include `numpy` and `scipy` for numerical linear algebra and PCA, `pandas` for data management, `cvxpy` for convex portfolio optimization, and `matplotlib` for visualization. Regime classification and covariance construction are implemented as modular functions to facilitate robustness checks.

A.2 Data Sources Summary

Table 3: Data Sources

Data Series	Source	Frequency	Period
U.S. Treasury CMT Yields (3m–30y)	FRED H.15 (DGS3MO–DGS30)	Daily	2005–2024
MOVE Index Proxy	FRED VIXCLS $\times$ 4.2	Daily	2005–2024
OTR/OFR Liquidity Proxy	FRED TEDRATE (SOFR post-2023)	Daily	2005–2024
Portfolio DV01 / KRDs	Computed from FRED yields	Daily	2005–2024

A.3 PCA Factor Loadings

Factor loadings $\boldsymbol{\Lambda}$ estimated from daily yield changes 2005–2024. Three factors retained ($K = 3$), explaining 95.42% of total variance. Values are unit-normalised eigenvectors.

Table 4: PCA Factor Loadings

Maturity	PC1 — Level (75.18%)	PC2 — Slope (13.90%)	PC3 — Curvature (6.34%)
3m	0.1028	−0.5601	0.5803
6m	0.1334	−0.4486	0.2230
1y	0.1928	−0.3945	−0.0748
2y	0.3111	−0.2561	−0.4112
3y	0.3563	−0.1627	−0.3586
5y	0.4023	−0.0004	−0.2072
7y	0.4102	0.1169	−0.0476
10y	0.3853	0.2008	0.1128
20y	0.3493	0.2938	0.3195
30y	0.3317	0.3116	0.3877

A.4 PCA Explained Variance by Component

Table 5: PCA Explained Variance

Component	Eigenvalue	Variance Explained	Cumulative
PC1 (Level)	195.40	75.18%	75.18%
PC2 (Slope)	36.13	13.90%	89.08%
PC3 (Curvature)	16.49	6.34%	95.42%

A.5 Regime-Conditional Annualized Factor Volatilities

Table 6: Annualized Factor Volatilities by Regime (bps)

Regime	Level bps	Slope bps	Curvature bps
NORMAL	145.2	59.6	43.0
STRESS	198.8	98.2	61.8
CRISIS	306.9	129.7	87.3

A.6 Isolation Forest Configuration and Regime Distribution

Table 7: Isolation Forest Parameters and Regime Distribution

Parameter	**Value**
Training window	2005-02-01 to 2007-12-31 (760 days)
Validation window	2007-07-02 to 2007-12-31 (131 days)
Out-of-sample period	2008-01-01 to 2024-12-31 (4,436 days)
$n_{\text{estimators}}$ (trees)	300
max_samples (subsample size)	256
Contamination	0.05
NORMAL regime days	3,840 (73.9%)
STRESS regime days	912 (17.6%)
CRISIS regime days	444 (8.5%)

A.7 Grace Rebalancing Threshold Parameters

Table 8: Grace Hedge Rebalancing Threshold and Cost Parameters

Parameter	**Value**	**Interpretation**
THRESH(NORMAL)	0.01	Rebalance if drift $>$ 1% of notional
THRESH(STRESS)	0.03	Rebalance if drift $>$ 3%; 3$\times$ less frequent than NORMAL
THRESH(CRISIS)	0.05	Rebalance if drift $>$ 5%; 5$\times$ less frequent than NORMAL
α	2.0	S_t amplification; at $S_t = 1$, threshold doubles within regime
κ	5.0	TC cost scaling constant (bps per unit DV01 turnover)
φ(NORMAL)	1.0	Baseline TC regime multiplier
φ(STRESS)	2.0	TC doubles in stress (bid-ask spread $\approx 2\times$ normal)
φ(CRISIS)	4.0	TC quadruples in crisis (bid-ask spread $\approx 4\times$ normal)