



SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

The Universality Premium in Treasury Yield Curve Risk

Rakshay Pawar

Senior Quantitative Finance Associate Codeflow Capital LLC, United States. ORCID 0009-0008-7367-4811,
E-mail: Rakshay@codeflowcapital.com

Abstract

This study develops a scope–efficiency–reliability framework for Treasury yield curve risk and quantifies the universality premium associated with protecting arbitrary linear key-rate exposures. Four principal prediction-set designs are compared over 6,502 walk-forward forecast dates from 2000 through 2025 using six US Treasury maturities. At the 99% nominal level, the Gaussian full-curve benchmark misses on 2.169% of dates; rolling and adaptive conformal full-curve sets reduce this to 1.000%. Broader coverage has a measurable set-size cost: the adaptive full-curve design requires a 57.1% larger mean standardised radius than calibration to a fixed family of five portfolios. The paired bootstrap 95% interval for the radius difference is [1.784, 1.944] standardised units. Full-curve conformal misses nevertheless remain temporally clustered, whereas the adaptive fixed-family design records 0.984% simultaneous misses and is not rejected by the full-sample conditional-coverage test. Adaptive updating increases radius relative to rolling calibration without resolving full-curve clustering in the base specification. The contribution is an explicit, empirically measured distinction between coverage scope, prediction-set size, and temporal reliability, giving practitioners a concrete basis for selecting and validating portfolio-aware risk bounds.

Keywords: conformal prediction, Treasury yield curve, distribution-free inference, key-rate duration, adaptive calibration, universality premium, fixed-income risk

1. Introduction

Treasury risk models must translate uncertainty about daily yield changes into bounds on portfolio losses. A Gaussian ellipsoid offers a convenient joint description, but its reliability depends on how well its covariance estimate and distributional assumptions represent observed curve shocks [11]. Regime shifts and stress episodes therefore motivate direct evaluation of both average breach frequency and the clustering of breaches over time.

Conformal prediction calibrates prediction sets from observed nonconformity scores [1, 2, 14]. Under exchangeability and the appropriate construction, finite-sample marginal coverage is at least the target level without a parametric distributional assumption. Treasury yield changes are dependent and may undergo distributional shifts, so that guarantee does not automatically transfer to the present series. Xu, Jiang and Xie [4] study ellipsoidal prediction sets for multivariate time series, while Gibbs and Candès [3] develop adaptive conformal inference under distribution shift. These methods motivate the score geometry and updating approach evaluated here.

The present study asks how those ideas interact with the scope of a risk-control mandate. A full-curve region can bound every linear key-rate exposure through its support function, whereas a score calibrated to a declared portfolio family targets a narrower event. The term-structure literature [12] and the conformal literature [2, 3, 4, 5] provide complementary foundations, but the practical comparison requires measuring set size and temporal reliability alongside average coverage.

Using 9,005 daily yield observations across six maturities, the analysis evaluates 6,502 out-of-sample dates. Its contribution is threefold: it measures the repair of Gaussian unconditional undercoverage through conformal calibration; quantifies a 57.1% radius premium for the evaluated full-curve scope; and distinguishes that coverage improvement from unresolved temporal clustering. These results make exposure scope an explicit model-design choice with observable consequences.

Section 2 reviews the conformal prediction framework and its fixed-income context. Section 3 describes the data and methodology. Section 4 presents the four main empirical results. Section 5 discusses practical implications and scope boundaries. Section 6 concludes.

2. Background and Related Work

2.1 Conformal Prediction

Conformal prediction links the rank of a new nonconformity score among calibration scores to a prediction set [1, 14]. In standard split conformal inference, held-out calibration observations determine a finite-

sample-corrected quantile. Under exchangeability, the resulting marginal coverage is at least the nominal target, with possible conservatism from discrete ranks. Angelopoulos and Bates [2] review this framework and its extensions. Distribution shift and serial dependence require additional assumptions, methods, or empirical validation; a distribution-free construction alone does not remove those difficulties.

2.2 Adaptive Conformal Inference

Gibbs and Candès [3] update an internal miscoverage level after each observation to target long-run coverage under distribution shift. A miss moves the state toward a more conservative subsequent threshold, while a non-miss moves it in the other direction. This study evaluates a bounded implementation with $\gamma = 0.10\alpha$, chosen in the research configuration. The clipping and finite calibration window are implementation choices; the original long-run guarantee is not treated as a guarantee of conditional validity for this modified dependent-data application.

2.3 Multivariate Ellipsoidal Sets

A scalar Mahalanobis score summarises a multivariate shock relative to its forecast covariance. The ellipsoidal approach is motivated by Xu, Jiang and Xie [4], whose multivariate time-series method has its own assumptions and conditional-coverage analysis. Here, a lagged EWMA covariance and a rolling empirical quantile provide a deliberately transparent specification. It is not a full implementation of their MultiDimSPCI algorithm; its coverage is evaluated directly through walk-forward backtesting.

2.4 Interest Rate Risk and Key-Rate Durations

Key-rate duration describes sensitivity to yield changes at selected maturity nodes [8]. Let x_t denote the six-dimensional yield change in basis points and w_p the portfolio's loss exposure in thousands of dollars per basis point. The linearised loss is $L_{t,p} = w_p^\top x_t$. This approximation connects curve regions to portfolio risk bounds [11]. Kupiec's test [6] assesses unconditional coverage; Christoffersen's tests [7] assess breach independence and joint conditional coverage. They are statistical diagnostics, not a complete regulatory-compliance assessment [17].

2.5 Research Gap

The empirical contribution is the joint evaluation of three distinct design questions: which exposures are covered, how large the resulting bounds are, and how breaches behave through time. We define the universality premium to measure the radius difference between full-curve and fixed-family calibration in the same Treasury dataset. This definition permits a direct comparison without assuming that either coverage scope is universally preferable.

3. Data and Methodology

3.1 Data

Daily US Treasury par yield curve rates are obtained from the US Department of the Treasury [13] at six constant maturities: 1, 2, 3, 5, 7, and 10 years. The cleaned sample contains 9,005 yield-level observations from 2 January 1990 through 31 December 2025, giving 9,004 one-day yield changes. The out-of-sample evaluation covers 6,502 forecast dates from 3 January 2000 through 31 December 2025. Every covariance forecast, calibration quantile, and adaptive state uses information available before the forecast date.

3.2 Covariance Estimation

The covariance state follows an exponentially weighted moving average with decay $\lambda = 0.97$, initialised from 500 daily shocks. Before each forecast it is shrunk 5% toward its own diagonal and receives a small numerical ridge. This fixed diagonal-shrinkage specification is distinct from an estimated Ledoit–Wolf shrinkage intensity [9]. With S_t denoting the lagged EWMA state, the forecast covariance is $\Sigma_t = 0.95S_t + 0.05\text{diag}(S_t) + \varepsilon_t I$. The ridge equals the larger of 10^{-10} and 10^{-10} times the mean diagonal of the shrunk matrix.

3.3 Full-Curve Prediction Sets

The Gaussian radius is the square root of the 99th percentile of a chi-square distribution with six degrees of freedom. Curve scores are $s_t = \sqrt{x_t^\top \Sigma_t^{-1} x_t}$. The rolling conformal radius is the order statistic at rank $\min\{n, \lceil (n+1)(1-\alpha) \rceil\}$ among the preceding $n = 750$ scores, with $\alpha = 0.01$ for the primary 99% analysis. The adaptive design uses the same history but replaces the target quantile level with a state updated by $\alpha_{t+1} = \text{clip}\{\alpha_t + \gamma(\alpha - I_t), 0.10\alpha, \min(0.25, 4\alpha)\}$, where I_t is the observed breach indicator. At 99%, $\gamma = 0.001$ and the state is bounded between 0.001 and 0.04.

3.4 Fixed-Family Prediction Sets

The fixed family contains three stylised bullet-like exposures and two curve trades. In the ordered 1-, 2-, 3-, 5-, 7-, and 10-year coordinates, their exposure vectors are (2, 15, 2, 0, 0, 0), (0, 2, 4, 37, 4, 1), (0, 0, 1, 4, 12, 68), (0, 50, 0, 0, 0, -50), and (0, 50, 0, -100, 0, 50), respectively, in thousands of dollars per basis point. These represent 2-, 5-, and 10-year bullet-like positions, a 2s10s steepener, and a 2–5–10 butterfly.

Portfolio losses are standardised by $\sigma_{t,p} = \sqrt{w_p^\top \Sigma_t w_p}$. The family score is the maximum standardised loss

across the five portfolios; its calibrated quantile gives the simultaneous bounds $q_t\sigma_{t,p}$. Adding a new exposure changes the calibration problem.

3.5 Backtesting Protocol

A full-curve breach occurs when the six-dimensional shock lies outside its ellipsoid; a fixed-family breach occurs when any declared portfolio loss exceeds its family-calibrated bound. Strict exceedance defines a breach. The support function of the full-curve ellipsoid gives the bound $q_t\sqrt{w^T\Sigma_t w}$ for any linear exposure w when the shock is inside that region. Kupiec and Christoffersen diagnostics are evaluated at 5% significance [6, 7]. Five prespecified regimes cover 2000–2006, 2007–2009, 2010–2019, 2020–2021, and 2022–2025. These are empirical backtests, not proofs of future or state-by-state conditional coverage.

3.6 Bootstrap Inference

Paired method differences use a circular block bootstrap with 5,000 replications and 21-day blocks [16, 18]. Standardised radius is dimensionless. Quantile loss is the pinball loss at the nominal quantile, first averaged over the five portfolios on each date and then across dates; it is reported in thousands-of-dollars loss-score units for the stated exposures. The three selected comparisons in Table 2 use pointwise 95% intervals. No stepdown adjustment for simultaneous inference [15] was applied.

4. Results

4.1 Coverage Rates

Table 1 reports the four principal designs at 99% nominal coverage. The Gaussian full-curve region records 141 misses, or 2.169% of dates, against approximately 65 expected at a 1% target. Rolling and adaptive full-curve conformal sets both record 65 misses, or 1.000% after rounding (65 divided by 6,502). The adaptive fixed-family set records 64 simultaneous misses, or 0.984%. These results establish a substantial empirical improvement in average coverage for the conformal designs.

Table 1. Full-sample coverage and standardised radius at the 99% level

Method	Misses	Miss rate (%)	UC p-value	CC p-value	Mean radius
Gaussian full curve	141	2.169	<0.001	<0.001	4.100
Rolling conformal full curve	65	1.000	0.9980	0.0166	5.085
Adaptive conformal full curve	65	1.000	0.9980	0.0166	5.126
Adaptive fixed family	64	0.984	0.8986	0.5302	3.263

Notes: 6,502 forecast dates, January 2000–December 2025. UC = Kupiec unconditional coverage; CC = Christoffersen joint conditional coverage. Curve methods are tested on full-curve containment; the fixed-family method is tested on simultaneous containment of the five declared portfolio losses. Non-rejection at 5% does not establish guaranteed conditional validity.

4.2 Conditional Coverage

The analysis identifies a distinction that average coverage alone would miss. Both full-curve conformal methods have an independence-test p-value of 0.0042 and a joint conditional-coverage p-value of 0.0166, indicating temporal dependence in their breaches. The adaptive fixed-family design is not rejected by the joint test ($p = 0.5302$). Its narrower score construction is associated with stronger full-sample backtest performance in this comparison, but that result does not establish conditional coverage at every date or in every regime.

4.3 Universality Premium

The adaptive full-curve radius averages 5.126, compared with 3.263 for the adaptive fixed-family radius. Their ratio implies a universality premium of 57.1%. The paired mean-radius difference is 1.863 standardised units, with a 95% bootstrap interval of [1.784, 1.944]. This quantifies the broader scope's additional radius under the evaluated design. It measures neither six-dimensional set volume nor economic or regulatory capital; those quantities would require separate definitions and analysis.

4.4 Adaptive vs. Rolling Calibration

Adaptive rather than rolling full-curve calibration adds 0.0414 standardised radius units (95% CI [0.0128, 0.0688]) and 0.1670 thousand-dollar quantile-loss units (95% CI [0.0822, 0.2559]). Both pointwise intervals exclude zero. Thus the base adaptive setting changes set size and the loss score without improving its 99% curve miss count or joint conditional-coverage result. This comparison establishes the value of evaluating adaptation as a design choice rather than assuming that updating alone improves every reliability metric.

Table 2. Selected paired differences at the 99% level

Comparison (left minus right)	Metric	Difference	95% CI
Adaptive vs rolling full curve	Standardised radius	+0.0414	[0.0128, 0.0688]
Adaptive vs rolling full curve	Quantile loss	+0.1670	[0.0822, 0.2559]
Adaptive full curve vs adaptive fixed family	Standardised radius	+1.863	[1.784, 1.944]

Paired circular block bootstrap: 5,000 replications and 21-day blocks. Intervals are pointwise 95% intervals. Standardised radius is dimensionless; quantile loss is in thousands-of-dollars loss-score units for the fixed exposure vectors, averaged across the five portfolios and dates. Positive differences indicate wider sets or higher loss scores.

4.5 Regime Analysis

Table 3 makes the local limits of full-sample performance visible. The adaptive fixed-family method records a 2.191% simultaneous miss rate in 2020–2021 despite non-rejection of its full-sample joint coverage test. Across the reported regimes, full-curve conformal miss rates are closer to the 1% target than the Gaussian benchmark's rates. No method is exactly at target in every regime. These results support evaluating average calibration, temporal dependence, and regime stability as distinct aspects of reliability.

Table 3. Coverage by prespecified market regime

Regime	Gaussian full curve	Rolling full curve	Adaptive full curve	Adaptive fixed family
2000–2006 Pre-crisis	1.944%	0.800%	0.800%	0.800%
2007–2009 Financial crisis	2.527%	1.596%	1.596%	1.197%
2010–2019 Low-rate period	2.079%	1.000%	0.960%	0.960%
2020–2021 Pandemic	3.586%	0.996%	1.195%	2.191%
2022–2025 Rate normalisation	1.804%	0.902%	0.902%	0.601%

All entries are miss rates at a 99% nominal coverage level. The first three columns of results measure full-curve breaches; the last measures simultaneous breaches within the declared five-portfolio family. The target rate is 1% for each method's own event. Regime sample sizes are 1,749, 752, 2,501, 502, and 998 dates, respectively.

5. Discussion

The universality premium gives exposure scope a measurable operational meaning. The evaluated full-curve design produces support-function bounds for arbitrary linear key-rate exposures, while fixed-family calibration targets five declared positions. Choosing the broader scope increases mean standardised radius by 57.1% in this sample. The measured difference provides a concrete input to model design, although its monetary significance depends on the exposures and decision rule to which the bounds are applied.

The unresolved clustering of full-curve misses also identifies a useful validation target. Changes in yield volatility may interact with lagged covariance estimates and historical score windows, and adaptive thresholds may respond only after a miss. The study does not separately identify those mechanisms. It instead establishes the empirical distinction: correcting average coverage in this setting is not sufficient to remove temporal dependence in breaches.

The narrower fixed-family score gives stronger full-sample backtest performance in the principal comparison, but its 2020–2021 deterioration prevents a universal reliability claim. A supplementary rolling fixed-family configuration likewise has a low average radius of 3.264 and a non-rejected joint coverage test ($p = 0.7642$). This supports interpreting the scope comparison separately from the choice to use adaptive rather than rolling calibration.

Several boundaries matter for interpretation. The study uses fitted Treasury par yields and linearised losses, excluding convexity, carry, financing, liquidity, optionality, and security-level execution effects. Treasury changed its curve-construction methodology in 2021. A supplementary historical-simulation benchmark is sensitive to boundary ties: at 99%, its five-portfolio simultaneous breach rate is 4.060% under strict exceedance and 4.752% under inclusive exceedance, with 77 tied dates. These are family-level diagnostics for individually calibrated historical bounds, rather than a claim that their family-wise target is 1%. Alternative covariance estimators, other sovereign markets, and different portfolio families could change the measured premium.

For practitioners, the contribution is a scope-aware validation framework. A stable, declared set of exposures can be evaluated with family-calibrated bounds, while changing exposures may justify the broader reusability of a full-curve region. Either choice should be accompanied by average-coverage, dependence, and regime diagnostics. The reported covariance, window, and learning-rate sensitivities reinforce this point: reliable coverage and compact sets are separate objectives, and the preferred design depends on the mandate being protected.

6. Conclusion

This study quantifies a universality premium in Treasury risk control and makes coverage scope an explicit, testable design choice. Across 6,502 walk-forward dates, conformal full-curve calibration reduces the Gaussian miss rate from 2.169% to 1.000% at a 1% nominal target. The adaptive full-curve region requires a 57.1% larger mean standardised radius than the evaluated five-portfolio family, with a paired radius-difference interval of [1.784, 1.944].

The analysis also establishes why average accuracy is only one component of reliability. Full-curve conformal breaches remain dependent (independence $p = 0.0042$; joint conditional-coverage $p = 0.0166$). The adaptive fixed-family set records 0.984% simultaneous misses and is not rejected by the full-sample joint test ($p = 0.5302$), while its pandemic-period breach rate reveals a local weakness. Adaptive full-curve updating adds 0.0414 radius units without resolving the base-case dependence result.

These findings provide a concrete contribution to fixed-income model validation: a measurable framework connecting exposure breadth, prediction-set efficiency, and temporal reliability. The empirical premium is a radius measure, and the backtests describe the tested sample and configurations. Within that scope, the framework helps distinguish a reusable curve-wide bound from a more focused portfolio-family bound and identifies the additional reliability checks needed for either design.

Extensions to nonlinear instruments, alternative covariance estimators, other sovereign curves, and changing exposure families can test how widely the framework applies. The present evidence supplies a transparent benchmark for that work: broader coverage has a measurable size premium, and closer average calibration can coexist with clustered misses.

References

- [1] Vovk, V., Gammerman, A., & Shafer, G. (2022). *Algorithmic Learning in a Random World* (2nd ed.). Springer.
- [2] Angelopoulos, A. N., & Bates, S. (2023). Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4), 494–591. <https://doi.org/10.1561/22000000101>
- [3] Gibbs, I., & Candès, E. (2021). Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34, 1660–1672.
- [4] Xu, C., Jiang, H., & Xie, Y. (2024). Conformal prediction for multi-dimensional time series by ellipsoidal sets. *Proceedings of the 41st International Conference on Machine Learning, Proceedings of Machine Learning Research*, 235, 55076–55099. <https://proceedings.mlr.press/v235/xu24m.html>
- [5] Tibshirani, R. J., Barber, R. F., Candès, E., & Ramdas, A. (2019). Conformal prediction under covariate shift. *Advances in Neural Information Processing Systems*, 32.
- [6] Kupiec, P. H. (1995). Techniques for verifying the accuracy of risk measurement models. *The Journal of Derivatives*, 3(2), 73–84. <https://doi.org/10.3905/jod.1995.407942>
- [7] Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review*, 39(4), 841–862.
- [8] Ho, T. S. Y. (1992). Key rate durations: Measures of interest rate risks. *The Journal of Fixed Income*, 2(2), 29–44.
- [9] Ledoit, O., & Wolf, M. (2004). Honey, I shrunk the sample covariance matrix. *The Journal of Portfolio Management*, 30(4), 110–119.
- [10] Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, 7(1), 77–91. <https://doi.org/10.1111/j.1540-6261.1952.tb01525.x>
- [11] McNeil, A. J., Frey, R., & Embrechts, P. (2015). *Quantitative Risk Management: Concepts, Techniques and Tools* (Revised ed.). Princeton University Press.

- [12] Diebold, F. X., & Li, C. (2006). Forecasting the term structure of government bond yields. *Journal of Econometrics*, 130(2), 337–364.
- [13] U.S. Department of the Treasury. (2026). Interest rate statistics: Daily treasury par yield curve rates.
<https://home.treasury.gov/policy-issues/financing-the-government/interest-rate-statistics>
- [14] Shafer, G., & Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9, 371–421.
- [15] Romano, J. P., & Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469), 94–108.
- [16] Politis, D. N., & Romano, J. P. (1992). A circular block-resampling procedure for stationary data. In R. LePage & L. Billard (Eds.), *Exploring the Limits of Bootstrap* (pp. 263–270). Wiley.
- [17] Basel Committee on Banking Supervision. (2019). Minimum capital requirements for market risk (BCBS d457). Bank for International Settlements. <https://www.bis.org/bcbs/publ/d457.pdf>
- [18] Lahiri, S. N. (2003). *Resampling Methods for Dependent Data*. Springer.