

A Two-Gate Test for Structural Edge in Systematic Trading Strategies

Sofiane Boutgajouft
OrgaX LLC
`contact@orgaxtech.com`

September 25, 2026

Abstract

We propose a distribution-free test of whether a systematic trading strategy generates a *structural* edge—one invariant across instruments, volatility regimes, and time—as opposed to regime-contingent returns, using only the trade-level R-multiple output and no access to the signal. Because a strategy with no edge can be distributionally invariant by construction, an invariance score alone is insufficient; we combine two orthogonal gates, one for the significance of the mean return and one for distributional invariance tested component by component against a Monte-Carlo null. We measure the edge on the raw directional return, which is free of any stop or target parameter, and show that judging it on a fixed reward-to-risk payoff instead would make the verdict depend on an arbitrary stop. The empirical application is built entirely on free, publicly available data: fourteen publicly specified trading strategies are reconstructed over 2008–2025, and none clears the edge gate. A proprietary FX strategy, regenerated on the same public price series and entering the test through its trade-level output alone, clears both gates ($t = 3.64$, $N = 582$). A placebo permutation, a graded sensitivity experiment and a Monte-Carlo power study calibrate the invariance gate and establish the sample size it requires. The framework is complementary to backtest-overfitting diagnostics and to factor attribution.

Keywords: structural edge; invariance test; Wasserstein distance; strategy validation; black-box testing; distribution-free methods

JEL: C12; C14; C58; G11; G12

1 Introduction

Allocators routinely face a strategy whose track record is observable but whose signal is not. A central question follows: is the strategy’s edge structural—invariant across instruments, volatility regimes, and time—or regime-contingent, concentrated in conditions that may not persist and prone to post-publication decay (McLean and Pontiff, 2016)? The distinction governs capital allocation,

yet it is typically assessed informally, by inspecting performance tables broken down by market regime.

This paper formalizes that practice into a calibrated, distribution-free test that operates on the trade-level output alone, with no access to the signal-generating mechanism. We treat the per-trade R -multiple distribution as the object of study and ask whether it is statistically invariant across three partitions: instrument, volatility regime, and time. The building blocks are established—Wasserstein distances for distributional comparison (Villani, 2009; Peyré and Cuturi, 2019) and Wasserstein-based goodness-of-fit testing with invariance reduction and calibrated finite-sample critical values (Hallin et al., 2021).

A single invariance score is insufficient. A zero-expectancy strategy that is i.i.d. across the partitions is, by construction, distributionally invariant, and a scalar invariance metric would misclassify it as structural. We therefore adopt two orthogonal gates. Gate 1 tests for a genuine edge; we measure it on the raw directional return, which carries no stop or target parameter, held to the credibility bar of Harvey et al. (2016). Gate 2 tests for invariance, component by component. A strategy is classified as structural only if it clears both gates.

Contribution. The statistical objects the test uses are not new. Its contribution is the setting it addresses: third-party due diligence on a strategy whose mechanism is not disclosed, where the only observable object is the realised trade sequence. Existing validation tools do not cover this case. Backtest-overfitting diagnostics (Bailey et al., 2017) and multiple-testing corrections (Harvey et al., 2016) both require the researcher’s trial history, which an external evaluator does not have and cannot recover from a single realised sequence. Factor attribution requires a return series aligned to factor dates and answers a different question: whether the return is compensation for known exposures, not whether the edge is invariant across the conditions in which it was earned. We assemble established components into a calibrated decision procedure for the signal-blind setting, and demonstrate that the procedure discriminates.

Reproducibility. The previous version of this paper drew its price data from a commercial terminal. The empirical section has been rebuilt on free public sources, so that every result reported here can be regenerated by any reader at no cost. Fourteen publicly specified strategies are reconstructed on daily OANDA price series, with policy rates from FRED and the volatility-regime partition from the CBOE VIX. The tested strategy has itself been regenerated on the same price series, so the comparison is like-for-like. The implementation is available at <https://github.com/OrgaxTech/two-gate-test>; the trade-level output of the tested strategy is provided as supplementary material. Its signal remains undisclosed, which is the setting the test addresses.

2 The Two-Gate Test

This section defines the test. Section 2.1 fixes the inputs and the information available to the tester; Sections 2.2 and 2.3 define the two gates in turn; Section 2.4 gives the null calibration; Section 2.5

states the decision rule. A worked example on one benchmark runs alongside the definitions.

2.1 Inputs and scope

The trade sequence. A strategy produces a finite sequence of trades indexed $i = 1, \dots, N$. Trade i has a realised profit-and-loss PnL_i and a risk amount $\text{Risk}_i > 0$ committed at entry, before the outcome is known. The *R-multiple* of trade i is the ratio

$$R_i = \frac{\text{PnL}_i}{\text{Risk}_i}, \quad (1)$$

so that $R_i = -1$ is a trade losing exactly the amount risked and $R_i = +3$ one gaining three times it. Expressing outcomes in units of risk committed at entry makes trades comparable across instruments, position sizes and volatility levels without any further normalisation. We write $\bar{R}$ and $\hat{\sigma}_R$ for the sample mean and standard deviation of $\{R_i\}_{i=1}^N$.

Partition labels. Each trade carries three labels, assigned from information available at its entry date:

- *P1, the cross-section:* the instrument traded. Its groups are the distinct instruments in the sequence.
- *P2, the volatility regime:* the prevailing level of market volatility at entry. We use the VIX on the entry date, split at its in-sample median, giving two groups (calm, stressed).
- *P3, time:* the position of the trade in the sample. We use equal blocks of trading days, giving three groups.

Write $d \in \{P1, P2, P3\}$ for a partition dimension, K_d for its number of groups, and $G_{d,1}, \dots, G_{d,K_d}$ for the index sets of the trades in each group. The regime and time labels depend only on the entry date, so no look-ahead enters their construction.

What the tester does not observe. The test consumes $\{R_i\}$ and the three label vectors, and nothing else. The signal that generated the trades, the entry and exit rules, the position sizing, and the parameters of the strategy are all unobserved. This restriction is not a simplification but the point of the construction: it is what allows a third party to validate a strategy whose mechanism the owner will not disclose.

Scope. The test applies to strategies defined by a signal producing discrete trades. A static position held throughout the sample has no entry or exit signal and therefore no trade structure to partition; such positions lie outside the scope of the test, and their return reflects a directional risk premium rather than a systematic edge.

2.2 Gate 1: is there an edge?

The statistic. Gate 1 asks whether the mean per-trade return is distinguishable from zero. The edge statistic is the one-sample t -statistic of the mean,

$$t_{\text{edge}} = \frac{\bar{R}\sqrt{N}}{\hat{\sigma}_R}, \quad (2)$$

and Gate 1 is passed if and only if $t_{\text{edge}} \geq t^*$ with $t^* = 3$, the bar advocated by Harvey et al. (2016) for claims of a new return predictor. We test the mean rather than the median: under a fixed-payoff representation the return distribution concentrates on a few mass points, and its median can be zero even when $\bar{R}$ is strongly positive.

Which return enters Gate 1. The quantity tested must not depend on parameters the analyst is free to choose. We therefore evaluate (2) on the *raw directional return* of each trade: the return of the instrument over a fixed holding window from entry, signed by the direction of the position. This carries no stop and no target.

The alternative—evaluating the edge on R-multiples produced by a fixed reward-to-risk payoff—is unsafe, because the verdict then depends on the stop distance rather than on the signal. Section 4 and the Online Appendix document the magnitude of this dependence: holding the signal fixed and varying only the stop multiple moves t_{edge} from strongly positive to strongly negative. The raw return has no such free parameter.

A strategy that reaches its own exits, as the tested strategy does, is a different case: its realised R-multiples are the outcome of its own logic, not of a payoff we imposed, so they enter Gate 1 directly.

2.3 Gate 2: is the edge invariant?

Gate 2 asks a different question: not whether the strategy earns a positive return on average, but whether the *distribution* of its returns is the same across the groups of each partition. A strategy whose edge is concentrated in one instrument, one volatility regime or one sub-period will show group distributions that differ; one whose edge is structural will not.

Comparing two groups. Let $\hat{F}_{d,j}$ denote the empirical distribution of the R-multiples of the trades in group j of dimension d . For two univariate distributions μ and ν , the Wasserstein-1 distance is

$$W_1(\mu, \nu) = \int_0^1 |F_\mu^{-1}(u) - F_\nu^{-1}(u)| du, \quad (3)$$

the mean absolute gap between their quantile functions (Villani, 2009; Peyré and Cuturi, 2019). We use W_1 rather than a moment comparison because it responds to any difference in shape, not only to a shift in mean: two groups with the same average outcome but different tails are distinguished by (3) and not by a t -test on their means.

Aggregating across groups. A dimension with K_d groups admits $K_d(K_d - 1)/2$ pairwise comparisons. The *structural invariance score* of dimension d is their average, divided by the dispersion of the whole sequence:

$$\text{SIS}_d = \frac{1}{\hat{\sigma}_R} \cdot \frac{2}{K_d(K_d - 1)} \sum_{j < k} W_1(\hat{F}_{d,j}, \hat{F}_{d,k}). \quad (4)$$

The normalisation by $\hat{\sigma}_R$ makes SIS_d scale-free: multiplying every R-multiple by a constant leaves it unchanged, so the score measures how far apart the group distributions are *relative to* the overall dispersion of the strategy's outcomes. A value near zero indicates group distributions that are nearly indistinguishable; larger values indicate divergence.

Worked example. Take the volatility dimension $P2$, which has $K_{P2} = 2$ groups and therefore a single pairwise comparison. For the tested strategy of Section 3, the calm-regime and stressed-regime R-multiple distributions are close: with $\hat{\sigma}_R = 1.271$, equation (4) gives $\text{SIS}_{P2} = 0.059$. For the 5-day mean-reversion benchmark the same computation gives $\text{SIS}_{P2} = 0.001$, and for the Bollinger reversal rule 0.135. Whether any of these values is small requires a reference point, which the next subsection supplies.

2.4 Null calibration

SIS_d is not zero under invariance. Finite groups drawn from the same distribution still differ by sampling noise, and the size of that noise depends on N , on K_d , and on how the trades are distributed across groups. A fixed threshold would therefore be miscalibrated: it would be too strict for a dimension with many groups and too permissive for one with few.

We calibrate each dimension against its own null. Under H_0 the R-multiples are i.i.d. across groups, so we simulate sequences of length N from a common distribution, assign them the *observed* partition structure, and recompute (4). Repeating $B = 10,000$ times yields the null distribution of SIS_d , and the threshold is its upper fifth percentile,

$$\tau_d = P_{95}(\text{SIS}_d \mid H_0). \quad (5)$$

Because SIS_d is scale-free, the simulation may use standard normal draws without loss of generality. Preserving the observed group sizes matters: τ_d grows with K_d and shrinks with N , so the thresholds of the three dimensions differ even within a single strategy.

Gate 2 is passed if and only if

$$\text{SIS}_d < \tau_d \quad \text{for all } d \in \{P1, P2, P3\}. \quad (6)$$

Testing component by component rather than aggregating into a single score is deliberate: it is what allows a failure to be attributed to a specific dimension.

On the interpretation of a pass. τ_d widens as N falls and as K_d rises, so the gate loses power in small samples and in dimensions with many groups. A component that passes should accordingly be read as an absence of detected deviation, not as established invariance. Appendix A reports the rejection rate of each component as a function of N , K_d and the size of the departure from invariance, and derives the sample size the test requires.

2.5 Classification

A strategy is classified as generating a *structural edge* if and only if it clears both gates:

$$t_{\text{edge}} \geq 3 \quad \text{and} \quad \text{SIS}_d < \tau_d \quad \forall d. \quad (7)$$

Both gates are necessary, and neither implies the other. A strategy with no edge whose returns are i.i.d. across the partitions is, by construction, distributionally invariant: an invariance score alone would classify it as structural. Conversely a strategy may earn a large mean return concentrated entirely in one regime, clearing Gate 1 while failing Gate 2. The two gates answer different questions, and the classification requires both.

When a strategy fails, the failing component locates the pathology:

Failure	Reading
$t_{\text{edge}} < 3$	No edge: the mean return is not distinguishable from zero
$\text{SIS}_{P2} \geq \tau_{P2}$	Regime-contingent: the edge depends on the volatility regime
$\text{SIS}_{P3} \geq \tau_{P3}$	Temporally unstable: the edge changes over the sample
$\text{SIS}_{P1} \geq \tau_{P1}$	Cross-sectionally unstable: the edge is instrument-specific

Gate 1 is evaluated first: an elevated invariance component is uninformative about a strategy that has no edge to begin with.

2.6 Relation to existing validation tools

The test addresses a setting the standard toolkit does not cover: an evaluator holds a strategy’s realised trade sequence and nothing else.

Backtest-overfitting diagnostics (Bailey et al., 2017) estimate the probability that an apparent edge arises from in-sample selection. They require the researcher’s trial history—the set of configurations examined—which an external evaluator does not have, and which cannot be recovered from a single realised sequence. The two-gate test is silent on selection bias and should be read alongside such diagnostics, not in place of them.

Multiple-testing corrections (Harvey et al., 2016) likewise require the number of trials to deflate the reported statistic. We borrow their $t \geq 3$ bar for Gate 1 but cannot apply the deflation itself, for the same reason.

Factor attribution regresses a return series on known factors and asks how much of the return they explain. It requires a return series aligned to factor dates rather than a trade sequence, and it

answers a different question: whether the return is compensation for known exposures, not whether the edge is invariant across the conditions in which it was earned. The two are complementary. In particular, the two-gate test characterises the robustness of an edge and does not decompose it into alpha and beta; a strategy carrying a large but stable factor exposure would clear both gates. Establishing that an edge is alpha in the asset-pricing sense requires a factor attribution, which we regard as a separate exercise.

3 Data and Strategies

3.1 Data

All price data are daily and free. Spot foreign-exchange series for seven major pairs (EURUSD, GBPUSD, AUDUSD, NZDUSD, USDJPY, USDCHF, EURGBP) are taken from OANDA via QuantConnect. Short-term policy rates, used to construct the carry benchmark, are taken from FRED. The volatility-regime partition ($P2$) uses the CBOE VIX, mapped to each trade’s entry date without look-ahead and split at its in-sample median of 17.49. The sample runs from January 2008 to December 2025; the first trade of the tested strategy occurs in February 2009, after its warm-up requirement is met.

No component of the empirical section requires a commercial data subscription. This is a deliberate change from the previous version and is what allows the full table below to be regenerated by a third party.

3.2 The tested strategy

The strategy under test is a proprietary systematic foreign-exchange strategy trading the seven pairs above. It identifies structural equilibrium levels by swing-point analysis and enters on displacement events around them; its parameters are proprietary and its signal is undisclosed. Only its realised R-multiple output enters the test, which is the setting of Section 2.1.

Its output has been regenerated on the OANDA price series used for the benchmarks, replacing the commercial-terminal series of the previous version. The regeneration also corrected a look-ahead bias in the implementation: stop and target resolution had been permitted on the same bar as entry, and is now deferred to the following bar. The corrected sequence contains $N = 582$ trades with mean R-multiple $\bar{R} = 0.192$ and $\hat{\sigma}_R = 1.271$. All figures reported below use the corrected implementation.

3.3 Benchmark strategies

To assess discriminant validity we reconstruct fourteen publicly specified trading strategies on the same instruments and the same sample. Each is defined by a transparent rule, replicable from published descriptions, and each uses its conventional published parameter setting; none is tuned on the sample.

- *Mean reversion.* 5-day reversal (fade the sign of the trailing 5-day return); Bollinger reversal (fade a two-standard-deviation deviation from the 20-day mean); RSI(2) reversal.
- *Trend.* 20/60 and 10/50 moving-average crossovers; MACD (sign of the 12/26 exponential moving-average gap).
- *Momentum.* 60-day time-series momentum; 12-month time-series momentum following Moskowitz et al. (2012); cross-sectional momentum (long the two strongest pairs, short the two weakest, on the trailing 60-day return).
- *Breakout.* Donchian range breaks at 20 and 55 days.
- *Carry.* Cross-sectional carry: long the two highest-yielding currencies and short the two lowest, rebalanced monthly, using published central-bank policy rates.
- *Reversal, cross-sectional.* The mirror of cross-sectional momentum on the trailing 5-day return.

Mapping signals to trades. Each signal is mapped to a sequence of non-overlapping trades. Entries sit on a fixed grid of twenty trading days, so the trade count does not depend on how quickly a rule exits and trades never overlap. For the invariance components, each entry is assigned a fixed reward-to-risk payoff: the risk unit is $R = c\hat{\sigma}^{(20)}$ with $c = 0.5$ and $\hat{\sigma}^{(20)}$ the trailing 20-day return volatility; a path reaching $+3R$ books $+3$, one reaching $-1R$ books -1 , and a 20-day time-stop books 0. Holding this payoff constant across benchmarks keeps the discretised shape of the R-multiple distribution comparable.

Crucially, this payoff is *not* used to judge edge. As Section 4 shows, the edge verdict under a fixed payoff depends on the arbitrary stop distance, so benchmark edge is measured on the raw directional return over the same twenty-day window. The tested strategy reaches its payoff through its own exit logic, so its realised R-multiples enter both gates.

A strategy excluded from the panel. An earlier version of this paper included a permanently long equity and gold position among the benchmarks. A static buy-and-hold position has no entry or exit signal and therefore no trade structure to partition, and it lies outside the scope stated in Section 2.1. We discuss the case in Section 4, since it bears on what the test does and does not establish, but it is not part of the benchmark panel.

4 Results

Table 1 reports the two-gate classification of the fourteen public benchmarks and of the tested strategy. Benchmark edge is measured on the pooled raw directional return; the tested strategy’s edge is measured on its realised R-multiples. Invariance components are computed on the fixed

reward-to-risk R-multiples for the benchmarks and on realised R-multiples for the tested strategy, with per-dimension null thresholds from $B = 10,000$ Monte-Carlo replications.

Table 1: Two-gate classification, 2008–2025. Edge for benchmarks is the t -statistic of the pooled raw directional return; for the tested strategy it is the t -statistic of realised R-multiples. $P1$: cross-section; $P2$: volatility regime; $P3$: time.

Strategy	N	t_{edge}	SIS_{P1}	SIS_{P2}	SIS_{P3}	Verdict
FX strategy (tested)	582	3.64	0.136	0.059	0.059	structural
Bollinger reversal	1195	2.95	0.086	0.135	0.060	no edge
60-day momentum	2529	2.24	0.054	0.121	0.076	no edge
XS momentum (60d)	1583	0.71	0.089	0.078	0.033	no edge
XS reversal (5d)	1846	0.57	0.059	0.022	0.059	no edge
20/60 trend	2529	0.53	0.044	0.045	0.067	no edge
MACD	2529	0.24	0.044	0.013	0.042	no edge
TS momentum (12m)	1967	0.02	0.057	0.111	0.066	no edge
Donchian 55	1461	−0.23	0.111	0.022	0.093	no edge
Carry (policy rates)	1153	−0.33	0.152	0.085	0.050	no edge
10/50 trend	2529	−0.33	0.052	0.018	0.014	no edge
RSI(2) reversal	1738	−0.56	0.070	0.037	0.043	no edge
5-day mean reversion	2528	−0.67	0.046	0.001	0.030	no edge
Donchian 20	2039	−1.01	0.091	0.021	0.029	no edge

Gate 1 separates the tested strategy from the public panel. On raw directional returns, no public benchmark attains the edge bar: t_{edge} ranges from -1.01 to $+2.95$, all below 3. The panel spans the standard families—mean reversion, trend, momentum, breakout and carry—at their conventional parameter settings, and the sample sizes are large, so the failures are not attributable to short samples. The tested strategy, by contrast, attains $t_{\text{edge}} = 3.64$ on the same price series.

The separation is not an artifact of the benchmark payoff. Measuring benchmark edge on the fixed reward-to-risk R-multiple rather than the raw return would make the verdict depend on the stop parameter: for a single mean-reversion rule, the same signal yields t_{edge} from $+18.7$ to -7.2 as the stop multiple varies from 0.5 to 1.5 standard deviations (Online Appendix). We therefore judge edge on the raw return, which has no free parameter, and reserve the fixed-payoff R-multiples for the invariance components.

Invariance alone is insufficient. Several benchmarks have low invariance components despite having no edge: the MACD rule has $\text{SIS}_{P2} = 0.013$ and the 5-day reversal $\text{SIS}_{P2} = 0.001$, both far below their thresholds. An invariance-only score would classify these unprofitable rules as structural. This is the empirical motivation for the edge gate: distributional invariance is necessary but not sufficient evidence of a structural edge.

Gate 2 discriminates among the failures. Three benchmarks show an elevated regime component relative to their own thresholds—Bollinger reversal, 60-day momentum and 12-month time-

series momentum—indicating volatility-regime dependence. These rules fail Gate 1 and are classified as having no edge, so the reading is diagnostic rather than decisive; but it shows the regime component responds where regime dependence is present.

The tested strategy clears both gates. With $t_{\text{edge}} = 3.64$ and all three invariance components below their $P_{95}(H_0)$ thresholds ($P1$: $0.136 < 0.240$; $P2$: $0.059 < 0.177$; $P3$: $0.059 < 0.184$), the tested strategy is classified as generating a structural edge. Its edge is not carried by a few outcomes: removing the ten largest R-multiples leaves $t_{\text{edge}} = 3.06$. Its mean R-multiple is positive in each of the three time blocks ($+0.130$, $+0.201$, $+0.239$), though it rises across them, a gradient the temporal component does not find sufficient to reject invariance.

What a passing verdict does and does not establish. The excluded buy-and-hold case makes the boundary concrete. A permanently long equity and gold position clears both gates on this sample: its mean return is significant and its return distribution is stable across instruments, regimes and sub-periods. It is nonetheless not an alpha-generating strategy but an exposure to a directional risk premium. The two gates measure whether an edge is present and whether it is invariant; they do not decompose it into compensation for known factor exposures and residual return. A strategy carrying a large but stable factor loading would clear both gates. Establishing that an edge is alpha in the asset-pricing sense requires a factor attribution, which is complementary to, and outside the scope of, a signal-blind test.

Sample size. The tested strategy’s $N = 582$ is marginally below the $N \geq 600$ this paper recommends for the temporal component (Appendix A). At $N = 582$ the regime component detects a ten-percentage-point win-rate gap with probability approximately 0.64 by interpolation of Table 2; the corresponding figure for the cross-sectional component is lower. The passing components should accordingly be read as an absence of detected deviation rather than as established invariance.

The invariance test is calibrated and sensitive. Two checks support the invariance gate. A placebo permutation of the tested strategy’s regime and time labels leaves its components at or below the permuted-null mean, confirming the low observed values reflect genuine invariance rather than an artifact of the construction. A graded experiment shows SIS_{P2} rising monotonically with an injected regime dependence, exceeding its threshold from a ten-percentage-point win-rate gap onward while remaining below threshold at a zero gap (Online Appendix).

5 Conclusion

We have proposed a distribution-free test of structural edge that operates on a strategy’s trade-level output alone, with no access to the signal. The test combines two orthogonal gates—the significance of the mean return and the distributional invariance of the output across instrument, regime, and

time—because invariance without an edge is not evidence of structure. We judge the edge on the raw directional return, which is free of discretionary exit parameters.

On fourteen publicly specified trading strategies reconstructed from free data over 2008–2025, no benchmark clears the edge gate. A proprietary FX strategy, regenerated on the same public price series and entering the test through its output alone, clears both gates. A placebo permutation, a graded experiment and a Monte-Carlo power study calibrate the invariance gate and establish the sample size it requires.

The test is intended as a complement to existing validation tools, not a replacement. It speaks to regime-robustness, not to in-sample selection bias, and should be read alongside backtest-overfitting diagnostics; it measures whether an edge is invariant, not whether it is compensation for known factor exposures, and should be read alongside a factor attribution. Its practical use is third-party due diligence on strategies whose mechanism is not disclosed, where the only available object is the realised output distribution. Extending the framework to a broader cross-section of audited live strategies, and exhibiting a live regime-contingent rejection at adequate sample size, are the natural next steps.

Data and code availability

The implementation of the test, the benchmark construction code and the power analysis are available at <https://github.com/SofianeBoutgajouft/two-gate-test.git>. All price and rate data are free: OANDA daily series via QuantConnect, FRED policy rates, CBOE VIX. The trade-level R-multiple sequence of the tested strategy is provided as supplementary material; its signal-generating code is not, by design, since the test requires only the output.

Disclosure of interest

The author is the founder of OrgaX LLC, which develops the proprietary systematic FX strategy used as one empirical illustration in this manuscript. This constitutes a financial interest in the strategy that appears in the empirical application. The methodological contribution of the paper—the two-gate test, the construction and analysis of the fourteen public benchmarks, and the calibration experiments—is independent of this interest and is reproducible by any third party on any trade-level output sequence.

Funding

No funding was received.

Use of generative AI

A generative AI assistant was used to support code development and manuscript editing. The author reviewed all content and takes full responsibility for the manuscript.

References

- Bailey, D.H., Borwein, J.M., López de Prado, M., Zhu, Q.J., 2017. The probability of backtest overfitting. *Journal of Computational Finance* 20, 39–69.
- Hallin, M., Mordant, G., Segers, J., 2021. Multivariate goodness-of-fit tests based on Wasserstein distance. *Electronic Journal of Statistics* 15, 1328–1371.
- Harvey, C.R., Liu, Y., Zhu, H., 2016. ... and the cross-section of expected returns. *Review of Financial Studies* 29, 5–68.
- McLean, R.D., Pontiff, J., 2016. Does academic research destroy stock return predictability? *Journal of Finance* 71, 5–32.
- Moskowitz, T.J., Ooi, Y.H., Pedersen, L.H., 2012. Time series momentum. *Journal of Financial Economics* 104, 228–250.
- Peyré, G., Cuturi, M., 2019. Computational optimal transport. *Foundations and Trends in Machine Learning* 11, 355–607.
- Villani, C., 2009. *Optimal Transport: Old and New*. Volume 338 of Grundlehren der mathematischen Wissenschaften. Springer.

A Power of the invariance gate

The recommendation $N \geq 600$ is derived rather than asserted. We simulate sequences with a controlled departure from invariance, apply the gate, and record the rejection rate. The design follows the graded experiment: R-multiples take the fixed-payoff values $\{+3, 0, -1\}$, and the departure is a win-rate gap between groups. The experiment is entirely synthetic, so the calibration does not depend on any particular sample.

Three readings follow. First, the type-I rate sits at or below the nominal 0.05 throughout, so the Monte-Carlo calibration of Section 2.4 is correctly sized and mildly conservative. Second, the regime component reaches conventional power against a ten-point win-rate gap between $N = 800$ and $N = 1000$; at $N = 600$ its power is 0.65. The recommendation $N \geq 600$ should therefore be read as the point at which the component detects a ten-point gap roughly two times in three, with $N \approx 1000$ required for the conventional 0.80.

Table 2: Rejection rate of the invariance component by sample size, number of groups and injected win-rate gap. A zero gap gives the type-I error rate, which should sit at or below the nominal 0.05.

Gap (pp)	$N=200$	300	400	500	600	800	1000	1200
<i>P2: two groups (volatility regime)</i>								
0	0.02	0.04	0.04	0.04	0.06	0.03	0.07	0.04
5	0.10	0.13	0.14	0.17	0.18	0.25	0.29	0.35
10	0.26	0.37	0.43	0.56	0.65	0.75	0.81	0.89
15	0.54	0.68	0.84	0.87	0.94	0.98	0.99	1.00
20	0.74	0.92	0.97	0.99	1.00	1.00	1.00	1.00
<i>P1: seven groups (instrument)</i>								
0	0.00	0.01	0.01	0.01	0.01	0.00	0.00	0.01
5	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
10	0.03	0.03	0.04	0.03	0.04	0.05	0.08	0.07
15	0.06	0.05	0.07	0.11	0.15	0.19	0.29	0.39
20	0.09	0.13	0.23	0.27	0.32	0.44	0.64	0.76

Third, and not documented in the previous version, power falls sharply with the number of groups. With seven groups the cross-sectional component reaches only 0.04 against a ten-point gap at $N = 600$, and requires a twenty-point gap and $N = 1200$ to approach 0.76. The mechanism is arithmetic: each group holds N/K_d observations, and the threshold must cover $K_d(K_d - 1)/2$ pairwise comparisons, so τ_d rises as the groups thin. A passing cross-sectional component at moderate N is consequently weak evidence, and should be reported as an absence of detected deviation rather than as established invariance. This asymmetry across partition dimensions is a limitation of the test as constructed.