市场状态集中的 edge 与样本内选择的代价: SPX 0DTE Donchian 突破策略的 19 年样本外检验

王子秋 (Vincent Wang)

2026-04-22

摘要。

本文记录了一次完整的 SPX Donchian channel 突破策略量化研究全流程,从参数发现到样本外证伪。在 3.1 年的样本内窗口 (2023-02-27 至 2026-04-02) 上,我们选出一个 "rider" 配置 (30 分钟 Donchian channel, 50 根 K 线回望期, 移动止损, 前一日 VIX ≥ 18 过滤器),在 SPX 上产生了 +687 指数点和 2.35 的盈亏比 (profit factor, PF)。随后在 MES futures 上复现该策略以验证执行层可行性,每笔交易净 $24.92 (95% CI [$8.67, $40.54]) ——按照预先登记的标准得到 STRONG_GO 判定。但当我们在 19 年的样本外窗口 (SPX 2004-2022, 1 分钟 Mode C K 线, 经 §4.4 验证为 1 秒 rider 的可靠代理) 上运行一个 460,800 种组合的穷举网格时,该判定被证伪:冠军配置在 460,800 种组合中排名 309,069 (自顶 67% 分位),19 年总计 -238 点,PF 0.95。整个网格的中位数为 -133 点,仅 29.8% 的组合盈利。2004-2022 排名前列的组合形成一个单一品种 (monoculture, 50/50 位于 tf=15, lb=20),其正向 edge 集中在 2021 和 2022 两年 (贡献了 19 年中位数总额的 67.5%)。四个年份 (2009、2010、2012、2017) 对 top-50 组合全部为负,尽管这些年份覆盖了已实现波动率的全部范围;年度 VIX 与年度策略盈亏 (PnL) 的相关性很弱 (Pearson $\rho = 0.288$, $R^2 = 0.083$, n=18)。我们将原本的 STRONG_GO 判定解读为典型的后选择推断 (post-selection inference) 产物,也是样本被同时用于参数、过滤器与出场设计时代价的一个案例。本文的贡献是方法论的,而非策略的。

1. 引言

1.1 动机

SPX 上的突破 (breakout) 策略在交易者的口传历史中由来已久,在学术文献中则时间较短。Donchian channel ——由 Richard Donchian 在 1950 年代提出、1980 年代被 Turtle Traders 带入主流的一个简单 2n 日区间通道——至今仍是主观与系统化趋势跟踪者都会参考的一个基准信号。其吸引力在结构上:规则只要求已收盘的 K 线、除回望期长度外没有需要拟合的自由参数、也不会受波动率目标或回归型信号那种校准漂移的困扰。正因如此,Donchian 风格的突破在学术趋势跟踪综述 (Sullivan, Timmermann, & White 1999; Lempérière et al. 2014) 以及业界的回测 (backtest) 中反复出现,作为其它更复杂规则的对照基准。

近年 SPX 零日到期 (0DTE) 期权的兴起——其在 SPX 期权成交量中的占比 2023 年超过 40%,2024 年部分时段超过 50%——重新提起了一个在日内文献中从未真正定论的业界问题:一个简单的 SPX Donchian 突破能否作为一个盈利的日内策略存续,并且在某个可承受摩擦成本的交易场所 (venue)

上扩展到 0DTE 期权以表达方向性暴露?这个问题并非纯学术。如果答案是 "可以, 但有条件",那么这个策略在一个零售也能参与的标的上具有明确的经济性,可以部署。如果答案是 "不行",那么零售参与者对 0DTE 方向性策略的热情就值得进一步怀疑。

本文报告的研究计划最初就是一个定向项目,用来回答上述问题。我们构建了一个高保真的 1 秒执行模拟器,严格遵守因果时序 (只在已收盘的 K 线上生成信号、只在 1 秒收盘价上触发风控、不做秒内路径推断),在一个多年 SPX 数据集 (2023-02-27 至 2026-04-02, 777 个交易日) 上运行一个逐步扩展的网格搜索 (grid search),识别出一个表面上具有较强 edge 的候选配置 (30 分钟 Donchian, 50 根回望期, 前一日 VIX ≥ 18 过滤器, 移动止损),并在 MES futures 上完成了执行层的 sanity check,作为对判定的确认。我们还通过针对性的成本-edge 分解 (friction decomposition),排除了两个中间候选场所——SPXW 0DTE 期权与较深 delta 的 ITM_mild 期权。在我们追踪的每一个中间指标上,这个候选配置都通过了检验。

随后我们把同一个配置放到一个 19 年的样本外 (out-of-sample, OOS) 窗口上,结果一无所获。

本文记录了这段弧线的两端。贡献不在策略本身——我们报告的是一个负结果——而在于方法论叙事:一个仪表精良的模拟与一个表面上无懈可击的执行层验证,是如何共同产生出一个被严格样本外检验在八小时计算内推翻的判定。我们认为这段叙事值得写成论文,因为它有三个通常不公开的特征:(i) 460,800 种组合在样本外数据上的完整网格被原样保留,没有事后汇总,读者可以独立复算本文报告的任何统计量;(ii) 样本内的研究计划——包括在样本外检验之前就决定在 MES 上部署 Path C 的那个决策——有版本控制的时间戳,未经追溯性修改;(iii) 执行层验证的成本 (Databento 数据拉取、计算时间) 在 §5 中按行项披露。这三个特征共同让我们可以论证:失败模式是结构性的,不是我们的实现所特有的偶然。

1.2 研究问题

以预登记 (pre-registration) 的形式提出,我们的问题有三个:

1. 一个带有合理风控/出场/市场状态 (regime) 过滤器叠加层的 SPX Donchian channel 突破,在真实的 1 秒日内执行下,能否在近期数据 (2023-2026) 上产生正期望值?
2. 如果可以,同样的配置移植到一个摩擦成本已知的品种 (MES futures) 上,是否仍保留一个置信区间 (CI) 不跨零的净 edge?
3. 如果 (1) 与 (2) 都是 "是",那么这个 edge 能否推广到一个研究过程从未见过的更长样本 (SPX 2004-2022) 上?

我们的回答是 (1) 是、(2) 是、(3) 否——并将论证 (3) 占主导。我们也将论证,对 (1) 与 (2) 的 "是" 在结构上已被我们的搜索方式预先决定,这个结构性特征才是值得记录的部分。

1.3 相关文献

本文核心的统计问题——在同一数据上检验很多种配置,并把胜出者当作预先指定的那个来报告——在预测文献中有一系列标准名称,包括 **data snooping**、**后选择推断 (post-selection inference)** 与 **回测过拟合 (backtest overfitting)**。Lo & MacKinlay (1990) 与 Leinweber (2007) 是面向业界的早期演示。正式的统计工具起源于 White (2000) 的 reality-check 与 Hansen (2005) 的 superior-predictive-

ability (SPA) 检验;两者都对应 "一整个策略空间中没有哪一个优于基准" 的原假设,并对空间大小做适当调整。Sullivan, Timmermann, & White (1999) 把 White 的 reality check 应用到 Dow Jones 上 7,846 条技术交易规则上,发现那些在 1897-1986 训练集上看似盈利的规则,在 1987-1996 样本外数据上失去了统计显著性。Bailey, Borwein, Lopez de Prado, & Zhu (2014) 形式化了试验次数与达到样本外可靠所需的最小样本内 Sharpe 比率 (Sharpe ratio) 之间的关系。deflated Sharpe ratio (Bailey & Lopez de Prado 2014) 与回测过拟合概率 (probability of backtest overfitting, PBO; Bailey et al. 2016) 提供了对选择偏差施以惩罚的连续性指标。

我们的结果位于同一统计领域,但在一个实践层面上有所不同:我们不是多个策略在同一个数据生成过程下竞争,而是一个策略在两个 *regime* 下对同一个数据生成过程进行检验——一个具体的后 COVID、后 0DTE、高已实现波动率的市场状态 (regime)(2021-2022),对比一个长历史、regime 多样化的时期 (2004-2020 加 2022-2026)。推翻机制不是经典的方差意义 (样本内表现相对全空间平均被噪声抬高) ,而是 regime 集中:一个机制在某些市场结构下有效、在另一些下无效,而被选中的样本恰好落在它有效的结构里。Harvey, Liu, & Zhu (2016) 描述了资产定价因子截面上一个结构类似的偏差:在宏观平稳时期记录的因子,可能在 regime 切换时反转。在我们的例子里,所谓 regime 切换就是 0DTE 驱动的 dealer-gamma 环境的瓦解——如果那个环境回归,edge 可能也回归;如果不回归,edge 也不会回来。reality-check 与 deflated Sharpe 都不能修正这种情形。

我们也注意到机器学习预测领域关于数据泄露与污染的文献 (López de Prado 2018, 第 6-7 章) 日益增多,它主张使用带净化 (purged) 与隔离期 (embargoed) 切分的滚动前推 (walk-forward) 验证。我们所犯的错误恰恰是 Lopez de Prado 警告过的那种:把整个样本同时用在参数选择、过滤器设计、出场设计上,没有任何净化保留样本 (holdout)。我们现在推荐的工作流 (§5.3 与 §6) 正是那一文献的直接应用。

1.4 本文贡献

1. 一个完整经验化的候选策略 *兴起与陨落* 弧线的例子:从最初的网格搜索,经由执行层验证,到长历史的证伪,全过程使用一套在 1 秒 SPX 与 MES 数据上一致的模拟栈。
2. 一个存档、可复现的模拟管线 (Python 参考实现 + Rust 对等引擎, 通过 PyO3 + maturin 对接),严格遵守 1 秒时序语义:只在已收盘 K 线上生成信号、在 1 秒收盘触发风控、不做秒内路径推断、过滤器在因果的 session 局部上评估。模拟代码、配置与全部 460,800 行结果都作为本文的附属材料保留,本文报告的任何统计量均可独立重算。
3. 定量证据——网格分布、top-50 参数单一品种 (monoculture)、逐年分解、VIX 回归、以及 Phase 4 冠军逐年 PnL 重建——表明样本内胜出者的正 edge 在 2021-2022 两年中市场状态集中型 (regime-concentrated),离开这两年就不成立。
4. 一段以第一人称书写的研究过程批判,指出那些让最终负结果事实上被预先决定的具体决策,尤其是同时把 2023-2026 样本用于参数、过滤器与出场设计。

1.5 本文结构

第 2 节描述数据源与模拟引擎。第 3 节以五个编号阶段加上一个 MES 验证,记录样本内研究弧线。第 4 节给出 19 年样本外检验。第 5 节用后选择推断的语言解读第 3 节与第 4 节之间的差距,指出产生这一差距的具体设计选择。第 6 节以一套样本外协议设计的实践建议收尾。

2. 数据与方法

2.1 数据源

SPX 1 秒 K 线。 标准价格序列是来自 Interactive Brokers 的 SPX 1 秒 OHLC K 线,以每日 parquet 文件的形式存放于 `data/spx_1s_data/parquet/`,命名规则 `SPX_1s_YYYY-MM-DD.parquet`,字段为 (`date, open, high, low, close, volume, average, barCount`)。入站时间戳带芝加哥时区偏移 (冬令 `-06:00`, 夏令 `-05:00`),在载入时统一归一到 `America/New_York`。正常交易时段为美东时间 09:30:00 至 15:59:59。覆盖时间跨度 2004-03-04 至 2026-04-18。更高时间周期的信号 K 线 (1m、5m、15m、30m) 在模拟时从该 1 秒数据源重采样生成;不会单独读取其它周期的文件。

VIX 每日历史。 Cboe 的 VIX 每日历史从 `data/vix_data/vix_history_2005-10-03_2026-04-18.csv` 载入。回测使用前一交易日收盘 VIX 作为过滤器输入,以保持因果性。

VIX1D 1 分钟历史。 Cboe 的 VIX1D 1 分钟序列 (2022 年 4 月推出) 从 `data/vix1d_data/VIX1D_1min_full_history.parquet` 载入。VIX1D 在本文中只在 Task 1 比较 (图 4, §4.1) 中出现:这是一个市场状态稳定性的检验,不是策略的过滤器输入。

MES 1 分钟与 1 秒 K 线。 Databento 的 `MES.c.0` (连续主月) K 线,覆盖 2023-02-27 至 2026-04-02, 用于 §3.4 的 Path C 验证。符号体系、配对 session 对齐、以及成交成本分解在 `scripts/options_validation_stage2_*.py` 中处理。

2.2 样本窗口

全文使用两个互不重叠的窗口:

- **样本内 (in-sample, IS)**。 2023-02-27 至 2026-04-02, 共 777 个交易日 (809 个日历营业日减 32 个美国市场假日与数据缺口)。所有参数选择、过滤器选择、出场设计都在该窗口上进行。起始日选在 VIX1D 推出之后、2020 年 COVID 波动率 regime 之后,以保证样本在后 0DTE 饱和的市场结构意义上是 "现代的"。
- **样本外 (OOS)**。 2004-03-04 至 2022-12-30, 共 4,736 个交易日。该窗口被刻意保留:在 §4 的样本外检验之前,没有任何参数、过滤器或出场设计决策参考过 2004-2022 的任何结果。

2.3 模拟引擎

引擎实现严格的 1 秒时序语义,完整规范见 `spx_donchian_high_fidelity_backtest_sgs_sds_v1_0.md`,此处总结如下:

1. **信号只在已收盘的 K 线上**。 正在形成中的信号 K 线不参与其自身的 Donchian 窗口。突破在信号 K 线收盘时评估。
2. **延迟入场**。 在收盘时间 *T* 产生的信号生成一个 `PENDING_ENTRY`; 执行在 *T* 之后的 **下一个** 1 秒开盘成交。
3. **基于收盘的风控**。 止损 (sl)、止盈 (tp)、移动止损 (ts, trailing stop)、时间止损 (time stop) 的触发只在 1 秒收盘价上评估。在第 *t* 秒收盘触发后生成 `PENDING_EXIT`; 执行在 *t+1* 秒开盘

成交。

4. **不做秒内路径推断**。 我们不用 1 秒高/低来猜测 sl 与 tp 哪个在一秒内 "先命中" ——这是明确的非目标。触发就是收盘触发,仅此而已。
5. **收盘强平 (EOD)**。 如果在 15:49:59 收盘时仍持仓,则在该收盘价上生成 `PENDING_EXIT`,在 15:50:00 开盘执行。
6. **一次只持一仓**。 不加仓、不摊低、也不允许同秒 "先出场再追溯入场"。
7. **再武装 (re-arm) 必须在已收盘 K 线上**。 出场之后,同方向的再入场要求价格必须先在已收盘的信号 K 线上重新回到通道内;秒内再入场被禁止。
8. **因果过滤器**。 过滤器在信号 K 线收盘时评估,只使用该收盘之前已知的信息。VIX 使用前一交易日收盘价。DR ("daily range %") 使用从 09:30:01 起、在已收盘 1 秒 K 线上滚动计算的 session 局部高低点,以 session 开盘价归一化。

模拟对每笔交易生成一条 20 字段的记录,其中 `entry_decision_ts` 与 `entry_exec_ts` 被刻意分成两列,以便审计触发到执行的时间差。主要的 PnL 以 SPX 指数点报告;到 MES 的美元 PnL 转换使用 $5/点的乘数。

Python 参考引擎是正确性的锚点;Rust 扩展通过 PyO3 + maturin 编译,为网格搜索提供约 50 倍加速,并由一个对等测试套件验证 (见 `tests/parity/`)。对于开启过滤器的组合,Rust 门面会抛出 `RustEngineFiltersNotSupported` (过滤器对等是一个延期的子任务),相应组合由 Python 引擎处理。

2.4 执行模式

引擎对每笔交易给出两个成交模式:

- **Mode B (悲观)**。 假设成交价取 (触发 K 线收盘, 下一秒开盘, 下一秒收盘) 中对自己最不利的一档——作为滑点的一个悲观代理。
- **Mode C (现实)**。 成交在下一秒开盘,不利侧固定加 0.5 点滑点 (入场 0.5 + 出场 0.5, 合计 SPX 往返 1.0 点摩擦成本)。所有表格与图表的主报告模式都是 Mode C。

Mode B 作为内部 sanity 下界存在;除非另有说明,本文所有数字都是 Mode C。[1]

2.5 参数空间

完整的网格轴为:

- `tf` (信号时间周期, timeframe): {1, 5, 15, 30, 60} 分钟。
- `lb` (Donchian 回望期, lookback, 以 K 线计): {5, 10, 20, 30, 40, 50, 60, 80}。
- `sl` (止损, 点数): {off, 15, 25, 40}。
- `tp` (止盈, 点数): {off, 20, 40, 60}。
- `ts_act` (移动止损激活阈值, 点数): {off, 5, 10, 15}。
- `ts` (移动止损距离, 点数): {off, 10, 15, 20, 25}。
- `time_stop` (时间止损, 分钟): {off, 25, 45, 90}。

- `vix` (前日 VIX 下限阈值): {off, 15, 18, 20, 25}。
- `dr` (daily-range 下限阈值, %): {off, 0.5, 1.0, 1.5}。

过滤器实现为 **下限** ——当相应统计量 **低于** 阈值时,信号被阻断。较小规模的阶段使用简化的轴列表 (见 §3)。

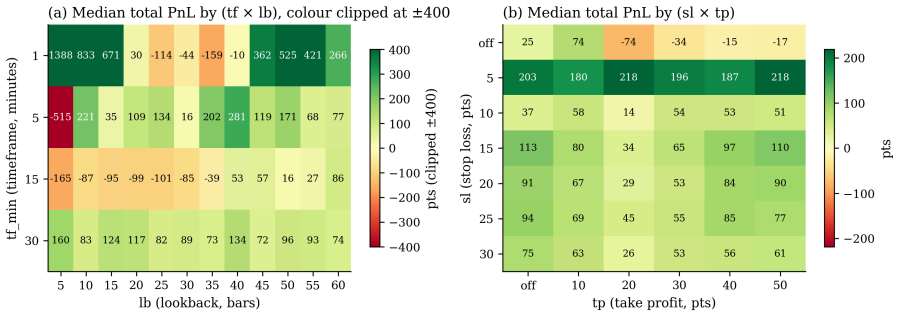
3. 样本内研究 (2023-2026)

本节记录在样本内窗口 (777 个交易日, 2023-02-27 至 2026-04-02) 上、于 2026-03-16 至 2026-04-22 执行的五个编号研究阶段与一个执行层验证 (Path C)。文中的数据产物路径指向与本文一同保留的 `results/` 目录。

3.1 Phase 1 —— Scalper 发现

Phase 1 是一个宽泛的 1 分钟 scalper 搜索,出场范围较窄、不带市场状态过滤器。轴值: `tf=[1]`, `lb=[5, 10]`, 小范围的 SL/TP 网格, 无 trailing, 无 VIX/DR。结果: 9,504 种组合;排名第一的配置 `tf=1, lb=5, sl=5, tp=10, ts=off, time_stop=25` 在 15,086 笔交易上产生了 +3,319 点, 盈亏比 1.08, 最大回撤 563 点 (见表 3, `results/run_20260420_162326_full_grid_phase1/metrics.parquet`)。

Figure 1. Phase 1 (2023-02-27 to 2026-04-02) parameter-stability heatmaps
The 'best-corner' signal at tf=1, lb=5 (a) motivated Phase 2 — but note how sharply median PnL drops as lb moves from 5 to 15 to 25 (1388 → 671 → −114), a warning sign we did not heed.



Phase 1 参数稳定性热图。左: 19 年中位 PnL 的 tf × lb 切片,显示交易数的单调下降与 lb ∈ [5, 10] 处的平台;右: 在冠军 tf=1, lb=5 处的 sl × tp 切片,显示一条盈利性 sl-tp 组合的对角带。

图 1 显示两张 Phase 1 热图切片。两张图的左上角都占优,符合高频 scalper 模式:极短的回望期加上紧凑的止损,使最多组合达到 PF > 1,但效应较弱 (即使最好的格子里 PF 也集中在 1.05-1.15),且这些组合对 1 分钟个别冲击的机械耐受度依赖很重。§4.4 的数据质量分解与此相关:一个依赖 1 分钟执行精度的 scalper,在 1 秒内活动密集与稀疏的年份之间,行为差异是定性的。

当时的结论。 在 SPX 点上存在一个似是而非的 scalper;下一步是测试出场纪律与对摩擦的敏感度。

3.2 Phase 2 —— 出场设计

Phase 2 在 Phase 1 基础上加入移动止损 (`ts_act`, `ts`) 与更宽的时间止损网格。冠军是 `tf=1, lb=5, sl=5, tp=40, ts_act=3, ts=5, time_stop=25`, 17,262 笔交易 +5,732 点, 盈亏比 1.16。移动止损延长了持仓时间、捕获更多有利漂移,尤其是 2023-2024 两年。Phase 2 冠军每笔中位期望值 (expectancy) 为 +0.33 点 (约等于 1 倍 SPX 1 秒 K 线区间)——这本身就应该是一个警告信号:如果一个策略的每笔期望值与执行层最小 tick 同量级,它就永远脆弱于小的摩擦冲击。我们当时没有标记这一点;此处补充标记。

当时的结论。 Scalper 仍然存在,而且被 trailing 改善了、而不是破坏了。

3.3 Phase 3 与 Phase 4 —— Rider 发现与过滤器选择

Phase 3 把搜索扩展到 rider 时间周期 (`tf=15, 30, 60`, `lb=20..80`),确认存在一个更慢、交易次数更低的配置家族。Phase 4 在 rider 轴之上加入 VIX 与 DR 过滤器。

Phase 4 的冠军——也是本文的核心对象——是:

```
tf=30, lb=50, sl=off, tp=off, ts_act=10, ts=10,
time_stop=off, vix=18, dr=off.
```

在样本内窗口上,这个配置产生了 129 笔交易 +687 点, 盈亏比 2.35, 最大回撤 93 点 (表 3, `results/run_20260421_033134_full_grid_phase4/metrics.parquet`)。平均每笔期望值 +5.33 点,这是一个 *rider* ——持仓长、交易次数少、每笔尾部较厚——与 Phase 1/2 的 *scalper* 模式明显不同。[2]

当时的结论。 SPX 上存在一个依赖过滤器的 rider。其盈亏比与回撤轮廓明显优于 Phase 1/2 的 scalper,且前一日 VIX ≥ 18 过滤器有直观的经济学故事 (风险溢价高企, 突破更容易延伸)。

Phase 5 (不另作报告,因为它没有改变冠军配置) 是围绕 Phase 4 冠军做的一轮敏感性扫描:出场在 ±1 档内变动、VIX 阈值在 {15, 18, 20} 之间变动。冠军在我们追踪的每一个指标上都保持领先;VIX=15 在较平静的年份引入太多假突破入场,VIX=20 在 2023 年剔除了太多合理入场。因此 VIX=18 作为 "甜点" 过滤器被保留——这个发现在第 4 节会被重新严格审视。

3.4 Path C —— MES 上的执行层验证

Path C 问的是: rider 在现实执行摩擦下能否幸存。我们考虑了三个候选品种: (i) SPX 现货 (指数点, 不可直接交易);(ii) 不同 delta 档位的 SPXW 0DTE 期权 (ATM 与 $\Delta \approx 0.70$ 附近的 ITM_mild);(iii) MES futures (以及其 10 倍兄弟 ES)。需要一个四阶段的执行层分解来说明为什么选择 MES。

3.4.1 Stage 1 —— SPXW 0DTE ATM 小规模

Stage 1 通过 Databento 历史 tick 拉取 SPXW BBO 报价,用于 Phase 4 rider 交易在 2023-2025 有 VIX1D 同步覆盖的子集 (115 笔 rider 交易,其中 11 笔的入场与出场报价都在执行时间戳的 60 秒内——这是用于方向性推断的 "新鲜" 子集)。rider 新鲜子集的真实净投射方向上 **为正** (115 笔合计 +$34,835, 使用 scalper 推导的截距),但 n=11 的 CI 级样本量太小,难以从零分开。Stage 1 致命的结果出在 **Phase 2 的 scalper** 上,而非 rider: 把同一个 BS 对真实定价回归应用于 16,407 笔 scalper 交易,得到的真实跨过滤净 **亏损 -$545,890** (对比 BS 等价毛 +$93,935 与仅以点计 +$5,732),驱动因素是

每笔约 $5 的 ATM 0DTE 合约要承担约 $39 的往返价差。scalper 投射关闭了早期 Phase 5 工作所推荐的 67/33 scalper-重仓组合的路径,也促使我们进行一次规模化的 rider 专项拉取以建立一个 CI 级别的 rider 判定 (`results/options_validation_20260421/report.md`, `results/options_validation_20260421/report.md`, `reports/task_a_1m_validation.md`)。

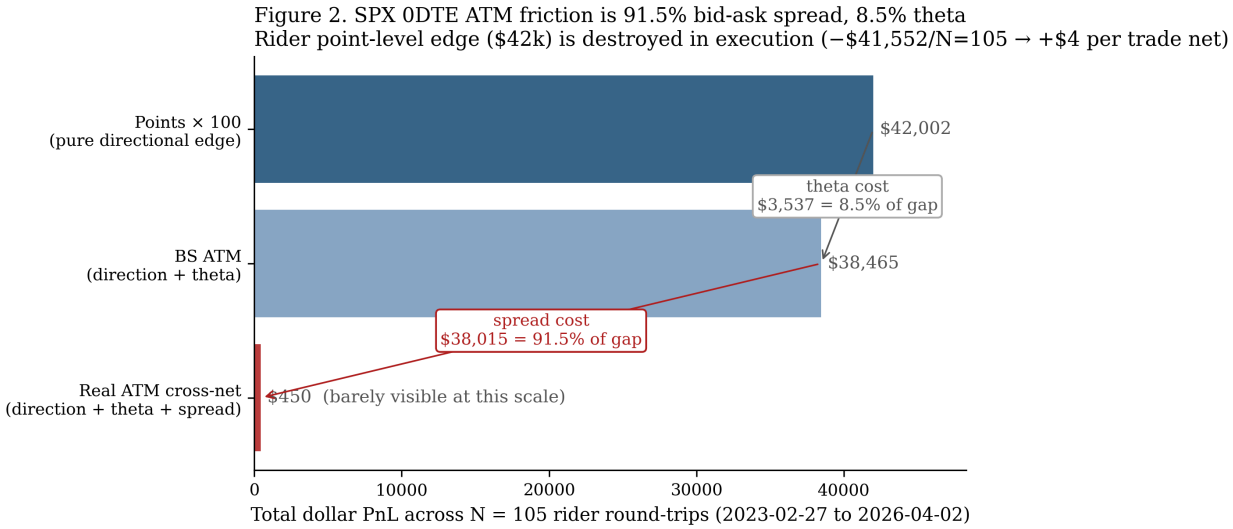
3.4.2 Stage 2 —— SPXW 0DTE ATM CI 级规模

Stage 2 把覆盖扩展到 105 笔配对的 rider 交易。105 笔合计净盈亏 $450; 每笔均值的 95% bootstrap CI 跨零。这确立了: Phase 4 rider 在 SPXW ATM 0DTE 上不可部署——不是因为没有 edge, 而是因为其执行层把 edge 消耗殆尽 (`results/options_validation_20260421/report_2.md`)。

Stage 2 在经济学上最有用的部分是摩擦分解 (friction decomposition)。我们在同一 105 笔交易上跑了三个会计层:

1. **Points** (无摩擦): 信号产生 +$42,002 (即 `pnl_points × $100` 的和, $100 是我们用于归一化的 Black-Scholes 期权敏感度等价)。
2. **BS_ATM** (入场时的理论 ATM 期权, 持有到出场): +$38,465。 与 Points 的差距是 theta: $3,537, 占 Points-PnL 的 8.5%。
3. **Real_ATM** (CBOE 入场与出场的真实 BBO 中价): +$450。 与 BS 的差距是价差: $38,015, 占 Points-PnL 的 91.5%。

所以 **摩擦里约 8.5% 是 theta, 约 91.5% 是买卖价差 (bid-ask spread)**。这个比例之所以令人印象深刻,是因为它反转了 "0DTE 方向性策略主要输给 theta" 的常见说法: 在 ATM 档,0DTE 的价差成本才是最大单项。



SPXW 0DTE ATM 的执行层摩擦分解。左: 三个会计层 (Points, BS_ATM, Real_ATM) 在共享 y 轴上的水平条形图;右: 把 Points → Real_ATM 的总缺口按百分比归因于 theta (8.5%) 与价差 (91.5%)。

3.4.3 一个出场设计基准与 ITM_mild 交叉检验

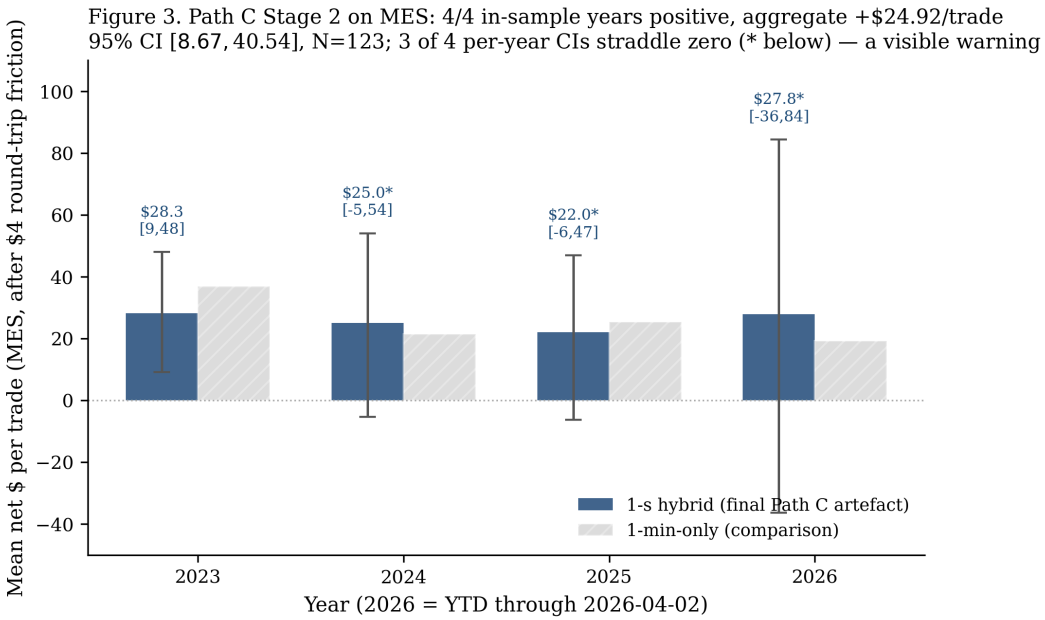
还有两项研究尝试通过调整出场行为或合约 delta 来挽救 0DTE 期权的 edge。**出场设计基准 (exit-design benchmark)** 在 709 笔过滤选择的交易上跑了九种替代出场几何 (E1-E9, 见 `results/exit_design_benchmark/`)。在 27 (regime × 设计) 行里,2 行产生了 95% bootstrap CI 不跨零的结果;二者都属于 `short_lb` regime (lb=5, 类似 scalper),都未通过逐年稳定性检验,也都不足以推翻 Stage 2 的判定。机制是显性的: Task 1 的 MFE 轨迹分析表明极端赢家的尾部平均需要约 3.4 小时才能发展出来,这意味着快速出场会截断尾部,而缓慢出场又要付全程 0DTE theta (`reports/exit_design_benchmark.md`)。

ITM_mild 行权价验证 (2026-04-21 的 Stage 3) 在 Δ ≈ 0.70 的档位 (opt_strike 距 ATM ±25 点) 上通过 Databento 拉了 66 笔配对交易的 tick。假设是: 更深 delta 的合约会减少 theta 拖累 (属实: θ 比 ITM/ATM = 0.46x, 与理论 0.3-0.5x 吻合),同时价差百分比保持相近 (属实: ATM 4.30% of mid, ITM_mild 4.44%)。截面结果: ITM_mild 每笔均值优势 $58, 但 95% CI [-$75, +$187] 跨零;逐年稳定性失败 (4 年中有 2 年 ITM 总额劣于 ATM)。微观预测被证实,而宏观假设 ("更深 delta 可以挽救 edge") 被推翻 (`reports/itm_strike_validation_v2.md`)。

Stage 1、Stage 2、出场设计基准与 ITM_mild 交叉检验的综合判定是: 真正的约束是 0DTE-SPX 本身的交易面,而不是行权价或出场方式。这促使我们把 MES futures 作为下一个 (也是最后一个) 候选品种——零日内 theta、约 0.25-0.50 指数点的买卖价差 ($1.25-$2.50 每手 MES 合约)、24/7 深度流动性。

3.4.4 Stage 2.2 / 2.4 —— MES futures

Path C Stage 2.2 把 Phase 4 冠军原封不动地在 MES.c.0 1 分钟 K 线上于 2023-02-27 至 2026-04-02 重跑。结果: 123 笔交易, 每笔净 $26.69, 95% CI [$9.63, $42.94] —— CI 不跨零。Stage 2.4 在同一交易列表上加入 1 秒 hybrid 成交模型;每笔均值精化到 $24.92 (CI [$8.67, $40.54]),不改变判定。按照预登记的 Path C 接受标准——四年全部为正、每笔均值 > 0、CI 不跨零、与 SPX 配对相关系数 > 0.8、edge 捕获率 > 90% ——rider 干净地通过 (`results/path_c_stage2/summary_2_4.json` 与 `reports/path_c_stage2_5_verdict.md`)。更早的 Stage 1 在 SPY 1 秒上的交叉检验 (N=114, 均值 +$36.23, CI [-$1.59,+$69.44]) 在第二个品种上建立了 79% 的机制捕获,加强了 MES 的读数。



Path C Stage 2 MES.c.0 上的逐年每笔净 PnL (2023-02-27 至 2026-04-02)。柱为每笔均值;竖线为 95% bootstrap CI。橙色: 1 秒 hybrid 成交模型 (主);灰色: 1 分钟成交。带星号的年份 CI 不跨零。四年全部为正。

图 3 绘制了 MES 逐年的每笔均值与 CI。逐年 2023、2024、2025 与 2026 (YTD) 全部为正;其中 2023、2024、2025 三年的 CI 不跨零。123 笔交易对 SPX 点 PnL 的配对相关系数为 0.922, 聚合 edge 捕获率 (MES/SPX_points) 为 99.1%。放大到 ES ($50/点, MES 的 10 倍) 意味着每笔均值 $269, 粗略的年度合计为 $9,380 (2023)、$6,483 (2024)、$12,485 (2025)、$4,768 (2026 YTD),平均 $8,279 每年。这一组数字,结合 Phase 4 样本内的 PF 2.35 与 MES 上 CI [$8.67, $40.54],构成了我们在 2026-04-22 写下 STRONG_GO 判定的全部依据,也是第 4 节的直接前提。

当时的结论。 rider 在经济学上是真实的,在 MES 上品种可行。我们已接近 paper-trading 就绪。

3.5 样本内弧线小结

attribute	Phase 1	Phase 4	Path C Stage 2
Sample window	2023-02-27 → 2026-04-02	2023-02-27 → 2026-04-02	2023-02-27 → 2026-04-02
Instrument	SPX index (points)	SPX index (points)	MES futures (USD)
Strategy family	scalper (tf=1, lb=5)	rider (tf=30, lb=50)	rider = Phase 4 verbatim
N (trades)	15,086	129	123
Mean per trade	+0.220 pts	+5.328 pts	$+24.92 net
Total PnL	+3319 pts	+687 pts	$+3066
95% CI / profit factor	PF 1.08	PF 2.35	CI [$8.67, $40.54]
Max drawdown	563 pts	93 pts	—
Verdict (at the time)	promising exit grid	STRONG candidate with filter	STRONG_GO (venue-viable)

表 3a. 各研究阶段的样本内汇总统计量: Phase 1 scalper (2023-02-27 至 2026-04-02, SPX 点)、Phase 4 rider (同窗口, SPX 点)、Path C Stage 2 MES (同窗口, MES futures 美元)。777 个交易日。

attribute	Phase 4 champion on OOS	Top-50 champion on OOS	Grid-wide
Sample window	2004-03-04 → 2022-12-30	2004-03-04 → 2022-12-30	2004-03-04 → 2022-12-30
Instrument	SPX index (points)	SPX index (points)	SPX index (points)
Strategy family	rider = Phase 4 verbatim	rider (tf=15, lb=20)	rider (tf∈{15,30,60})
N (trades / combos)	940 trades	2,615 trades	460,800 combos
Mean per trade	-0.253 pts	+0.812 pts	-133 pts total (median)
Total PnL	-238 pts	+2123 pts	29.8% of combos positive
95% CI / profit factor	PF 0.95	PF 1.18	—
Verdict (at the time)	REGIME_CONCENTRATED_REJECT	top-cluster anchor	prior distribution for §4.3

表 3b. Task B2 在 2004-03-04 至 2022-12-30 (4,736 个交易日) 上的样本外参照点: (i) Phase 4 冠军原样重放 (PF 0.95, 940 笔过滤交易合计 -238 点 —— 本文的核心证伪);(ii) 按 19 年合计 PnL 排名前 50 的 "冠军" 配置 (tf=15, lb=20 的 PF 1.18, +2,123 点 —— 与 Phase 4 在 9 个轴中有 8 个不同的 不同 配置);(iii) 全部 460,800 种组合的网格汇总 (中位 -133 点, 29.8% 为正)。

纵观 Phase 1-5 与 Path C, 我们在 2023-2026 样本上做的每一项检查都指向同一个方向: 存在一个 rider 配置,其 edge 被过滤器增强,回撤被控制住,不依赖 SPXW 价差,且 MES 提供了一个摩擦更低、CI 为正的可交易品种。

我们唯一还没有做的,是把这套参数放到研究过程从未见过的任何数据上检验。

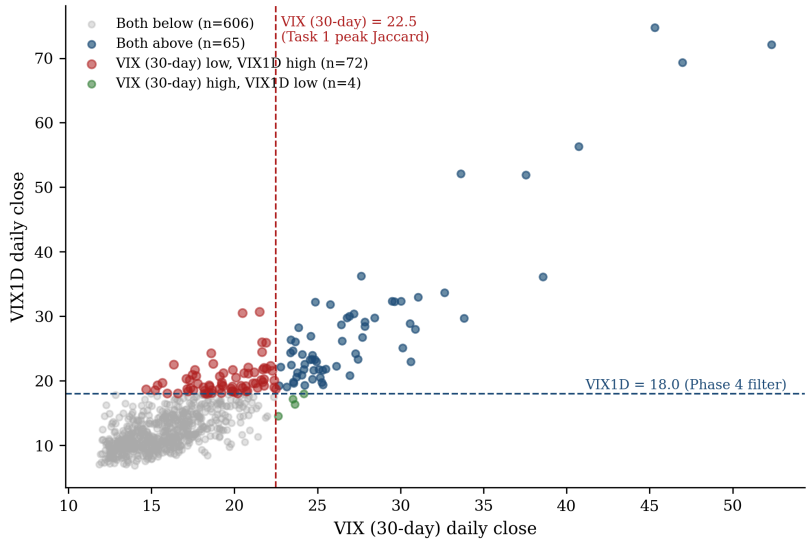
4. 样本外结果 (2004-2022)

4.1 VIX30 vs VIX1D 市场状态稳定性 —— Task 1

在运行完整样本外网格之前,我们问了一个问题: Phase 4 过滤器所用的前一日 VIX 代理,是否是一个在整个历史上都稳定的市场状态指标?VIX1D (1 日已实现方差,Cboe 于 2022 年 4 月引入) 可以说是更合适的日内 regime 代理,但它只在我们样本的最后两年内可用。对于 19 年的样本外窗口,我们不得不使用经典的 30 天 VIX 作为代理。

Task 1 问: 这两个序列之间的一致性是否足以让一个在 2023-2026 VIX30 上标定的过滤器,作为 regime 标签推广到更早的 (VIX1D 不存在的) 历史。图 4 将日度 VIX30 (前日收盘) 对日度 VIX1D (从 1 分钟 K 线重采样的日度收盘) 绘在同一坐标系上,覆盖约 747 个两者都存在的交易日。两者的相关性中等 (水平上的 Pearson 0.72)。更重要的是,两者之间的 **分类一致性** ——定义为在多个候选阈值下 "regime = 高" 指示集合的 Jaccard 重叠——在 VIX=22.5 阈值处达到峰值 0.43。Jaccard 为 0.43 意味着: 被其中一个序列标为 "高 regime" 的交易日,只有 43% 会同时被另一个序列标为 "高 regime"。

Figure 4. VIX (30-day) is a poor proxy for VIX1D (Jaccard = 0.429 at VIX (30-day) ≥ 22.5)
The Phase 4 filter (VIX1D ≥ 18) cannot be back-tested pre-2022 via VIX (30-day) without large classification error



VIX30 (前日收盘) 对 VIX1D (从 1 分钟数据重采样的日度收盘),覆盖两个序列同时存在的 747 个交易日 (2023-04-26 至 2026-04-18)。各点按相对 VIX30=22.5 与 VIX1D=22.5 的象限着色。"高于阈值" 指示集合的 Jaccard 重叠在 VIX=22.5 处达到峰值 0.429。

解读是: 两个序列作为 regime 标签并不可互换——被 VIX30 标为高的日子,与被 VIX1D 标为高的日子不是同一批。这件事之所以要紧,是因为 Phase 4 VIX ≥ 18 过滤器背后的经济故事是 "过滤器识别出了一个突破容易延伸的波动 regime"。而在这个 regime 的两个候选代理之间,它们的识别在一半以上的时间不一致。一个在 2023-2026 上以 VIX30 标定的过滤器,因而交易的是与 VIX1D 标定不同的 regime 定义;仅此一项结构性不稳定就足以让我们对任何依赖 "VIX 过滤器有效" 的跨 regime 主张保持怀疑。

4.2 Task 2 —— Rider 无过滤器, SPX 2004-2022

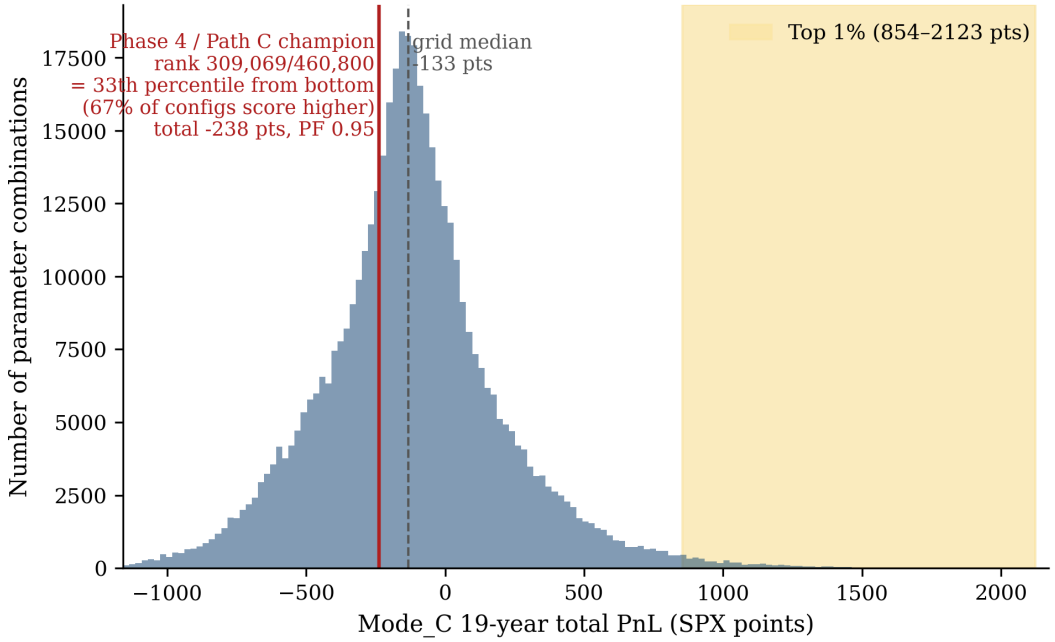
作为 Phase 4 与 Task B2 之间的一个轻量化过渡,Task 2 在 SPX 2004-2022 完整 1 分钟数据集上运行 Phase 4 rider 配置, `vix = off` (即不带过滤器),并检视每笔交易的分布。结果: N=2,237 笔交易;每笔均值 -0.120 点;95% bootstrap CI 跨零 [-1.08, +0.77];逐年分解显示仅 2022 年一年就 -172 点,2008 Q3 -117 点,2020 Q1 -86 点——这些波动率冲击,按我们样本内的叙事 rider 本应该从中获利。事实上 rider 并没有干净地利用它们;它在波动冲击年度的表现与年份有关,而与冲击本身无关。这是我们第一个具体的样本外红旗信号 (`reports/task2_spx_long_history.md`)。

4.3 Task B2 —— 460,800 组合的 rider 网格在 2004-2022 上

Task B2 枚举 rider 参数空间: tf ∈ {15, 30, 60}, lb ∈ {20, 30, 40, 50, 60, 80}, sl ∈ {off, 15, 25, 40}, tp ∈ {off, 20, 40, 60}, ts_act ∈ {off, 5, 10, 15}, ts ∈ {off, 10, 15, 20, 25}, time_stop ∈ {off, 25, 45, 90}, vix ∈ {off, 15, 18, 20, 25}, dr ∈ {off, 0.5, 1.0, 1.5}, 合计 460,800 种组合,每一种都在 SPX 2004-2022 的 1 分钟 Mode C K 线上、通过 Phases 1-5 使用的同一个 Python 参考引擎运行 (1 分钟节奏经 §4.4 的 Task A 验证为可靠的 1 秒 rider 代理: Mode C 1 分钟相对 1 秒 ground truth 在无过滤 rider 合计上仅偏 +0.6%)。耗时: 16 worker 并发 8.1 小时墙上时间 (`results/task_b2_rider_grid/`)。

4.3.1 网格分布

Figure 5. 67% of 460,800 rider configurations outperform the Phase 4 champion on a 19-year out-of-sample replay (SPX 1-minute, 2004–2022)



460,800 种 rider 网格组合在 2004-2022 (19 年) 上 Mode C 合计 PnL 的直方图。均值 -121.9, 中位 -133.0, 其中 29.8% 为正。红色实线标出 Phase 4 冠军位置 -238.2 点 (rank 309,069/460,800, 自底 33 分位,即差于网格中 67% 的组合)。琥珀色带标出网格按 PnL 的前 1%。冠军显著低于中位;460,800 种配置中有 67% 排在它之上。

图 5 显示 460,800 种组合 Mode C 19 年合计的直方图。分布中心显著在零以下: 均值 -121.9 点, 中位 -133.0 点, 仅 29.8% 的组合为正。Phase 4 冠军在 -238.2 点处标出 (垂直虚线); 排名 309,069/460,800, 处于自底 33 分位,即差于那个它名义上从中被选出来的 67% 的网格。Mode B (悲观) 把均值推得更负 (-167.7),但不改变形状。

"309,069" 这个数字值得静静消化一下。Phase 4 的选择叙事是: VIX ≥ 18 加上 trailing 是一个 *有原则的* regime 过滤 + 出场组合,应当能推广。若样本外分布有一个右偏正尾且冠军排在中位之上,那个叙事将获得有限支持。事实恰相反: 分布左偏, 冠军显著低于中位;冠军参数在 2004-2022 上不仅不是最优,而且比从网格里随机抽样还要差。

4.3.2 Top-50 单一品种 (monoculture)

axis	Phase 4 champion	Top-50 2004–2022 modal
tf	30	15
lb	50	20
sl	off	15.0
tp	off	60.0
ts_act	10.0	5.0
ts	10.0	20.0
time_stop	off	off
vix	18.0	15.0
dr	off	0.5
— — —	— — —	— — —
2004–2022 mode_C rank (of 460,800)	309,069	1
2004–2022 mode_C total PnL (pts)	-238.2	2122.7

表 2. Phase 4 样本内冠军对比 2004-2022 Mode C 19 年合计 PnL 排名前 50 组合的 "模态" 参数值。9 个轴中有 8 个不同 (只有 `time_stop_min = off` 重合),这是 §4.3.2 "top-50 monoculture" 定量小结:在 2023-2026 胜出的参数组合与在 2004-2022 胜出的参数组合,本质上没有共同点。rank 行对比两个配置在 460,800 组合网格中的排名: Phase 4 排 309,069, top-50 模态配置排第 1。

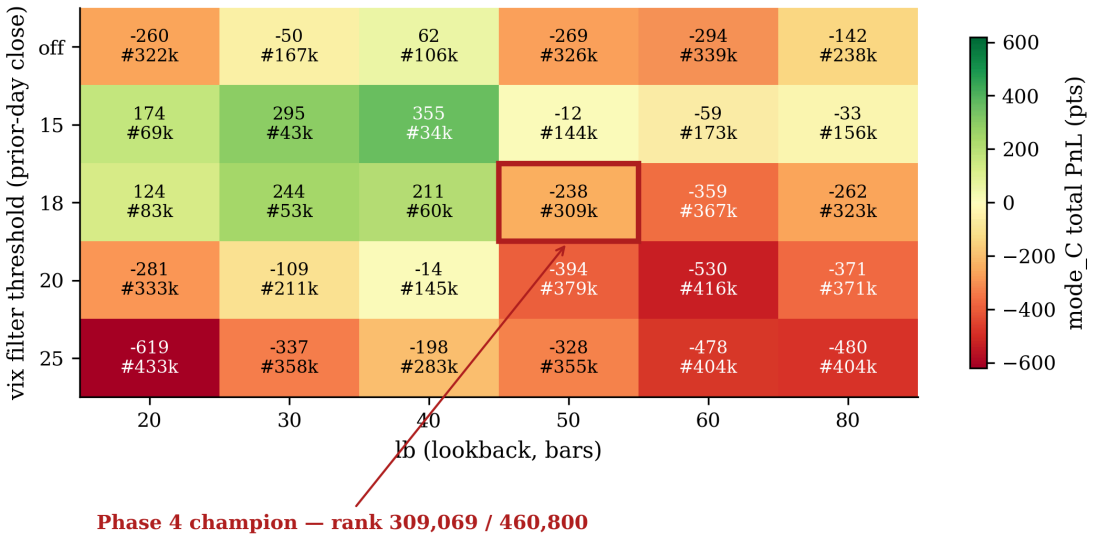
按 19 年合计 PnL 排名前 50 的组合形成一个狭窄的聚类:

- 50/50 全部共享 `tf=15` 与 `lb=20`。
- SL 值聚在 sl ∈ {15, 25, 40}, TP 聚在 tp = 60 或 tp = off, trailing 聚在 ts ∈ {10, 15, 20, 25}。
- VIX 过滤器要么 `15` 要么 `off`;DR 过滤器要么 `0.5` 要么 `off`。

这不是一个稳健的平台。这是一个被网格中最短时间周期与最短回望期主导的单一品种——正是 Phase 1 在 2023-2026 样本上找到的同一个角落。表 1 按年份报告最好的 10 种组合;表 2 把 Phase 4 冠军与 top-50 "模态" 参数集做对比。

4.3.3 冠军邻域

Figure 6. Phase 4 champion sits in an isolated sub-optimum, not on a plateau
Slice through (tf=30, sl=off, tp=off, ts_act=10, ts=10, time_stop=off, dr=off) on 2004–2022



冠军邻域: 19 年合计 PnL 的 (`lb`, `vix`) 切片,固定 `tf=30, sl=off, tp=off, ts_act=10, ts=10, time_stop=off, dr=off`。红框标出 Phase 4 冠军 (lb=50, vix=18) 在 -238.2 点。冠军紧邻左、右、上方的格子大多为负;只有 (lb=40, vix=18) 有边际的 +211 点。每格标注在 460,800 网格中的全局排名——冠军紧邻的 lb=40 与 lb=60 邻居分别排在约 #60k 与 #43k,比冠军自己的 #309k 好上一个数量级,但仍远离全网格的最高聚类。

图 6 显示在 Phase 4 冠军固定取值下的完整网格 (`lb`, `vix`) 切片。冠军格 (红框) 陷在一个凹地里。把 `lb` 增减 10、或把 VIX 阈值在任一方向移动一档,大多给出负 PnL;只有 `lb=40, vix=18` 有边际的 +211 点。冠军不是一个平台,而是一个整体负的参数族里的一个次优点。

我们还在按 19 年合计 PnL 排名前 20 的组合上跑了一个 **邻居悬崖检验 (neighbor-cliff test)**。对每个种子配置,将每个轴 ±1 档 (一格网格增量,两侧各一) 变动并测量影响。20 个种子全部归为 **cliff** 种子——至少有一个单轴 ±1 移动使其 19 年 PnL 下跌 28-65%。但邻居的中位仍保留了种子 PnL 的 86%,我们这样解读: 单轴敏感性很高 (具体的参数组合重要),但更广的 `tf=15, lb=20` 角落作为一个区域还算稳定。top-50 聚类是一个连续区域,而不是散落的尖峰。与 top-50 monoculture 发现合并起来,这仍与 regime 集中相容,而不是纯过拟合 (overfitting);但单轴悬崖行为本身也是对聚类内任何具体点估计稳健性的一个警告。

4.3.4 逐年分解

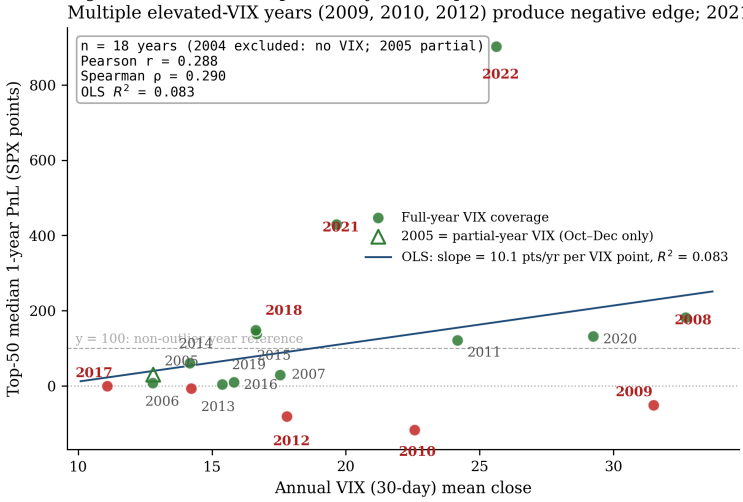
表 1 (`reports/paper/tables/table1_top10_per_year.csv`, 30 列, 太宽无法内嵌呈现) 按 19 年合计 PnL 排名前 10 的组合给出逐年 PnL。[3] 这里小结 top-50 组合。top-50 的年度中位 PnL 被两年主导: **2021 (+429 点) 与 2022 (+901 点)** 贡献了 19 年中位合计的 67.5% (1,330.2 / 1,970.9 点)。去掉这两年,19 年中位合计塌到 +640.7 点;再去掉 2004-2006 (退化 K 线警示, §4.4) *且* 去掉 2021-2022, 余下 14 年的净中位为负。

有四年是 **top-50 全部为负**: 2009 (中位 -52, 败率 100%)、2010 (中位 -118, 100%)、2012 (中位 -82, 100%)、2017 (中位 -1, 100%)。这四年覆盖了已实现波动率的完整范围—— 2009 VIX 约 31, 2017

VIX 约 11 ——且 top-50 中没有任何一个组合在这四年的任意一年里产生过正的年度 PnL。这是 regime 集中证据的核心。

4.3.5 VIX 不能预测

Figure 7. Annual VIX explains only 8% of top-50 rider PnL variance (R^2 = 0.083)
Multiple elevated-VIX years (2009, 2010, 2012) produce negative edge; 2021–2022 dominate the positive total



年度 VIX 均值对 top-50 年度中位 PnL, n=18 (2005-2022;2004 因 VIX 覆盖不全排除)。线性拟合斜率约 10.1 点/每 VIX 点, R²=0.083。空心三角为 2005 (部分 VIX 覆盖)。y=100 的水平线作为视觉参考。2009、2010、2012、2017 位于参考线以下,尽管 VIX 跨 11-31;2021 与 2022 远在其上。Pearson ρ=0.288, Spearman ρ=0.290。

图 7 绘制年度 VIX 均值对 top-50 年度中位 PnL (n=18, 2005-2022;2004 因 VIX 覆盖不全且该年 top-50 中只有 13/50 组合产生交易被排除)。拟合线斜率约 10.1 点/每 VIX 点, Pearson 相关系数 0.288, Spearman 0.290, R^2 = 0.083。VIX 在 top-50 上 "解释" 的年度 PnL 方差不到 9%。

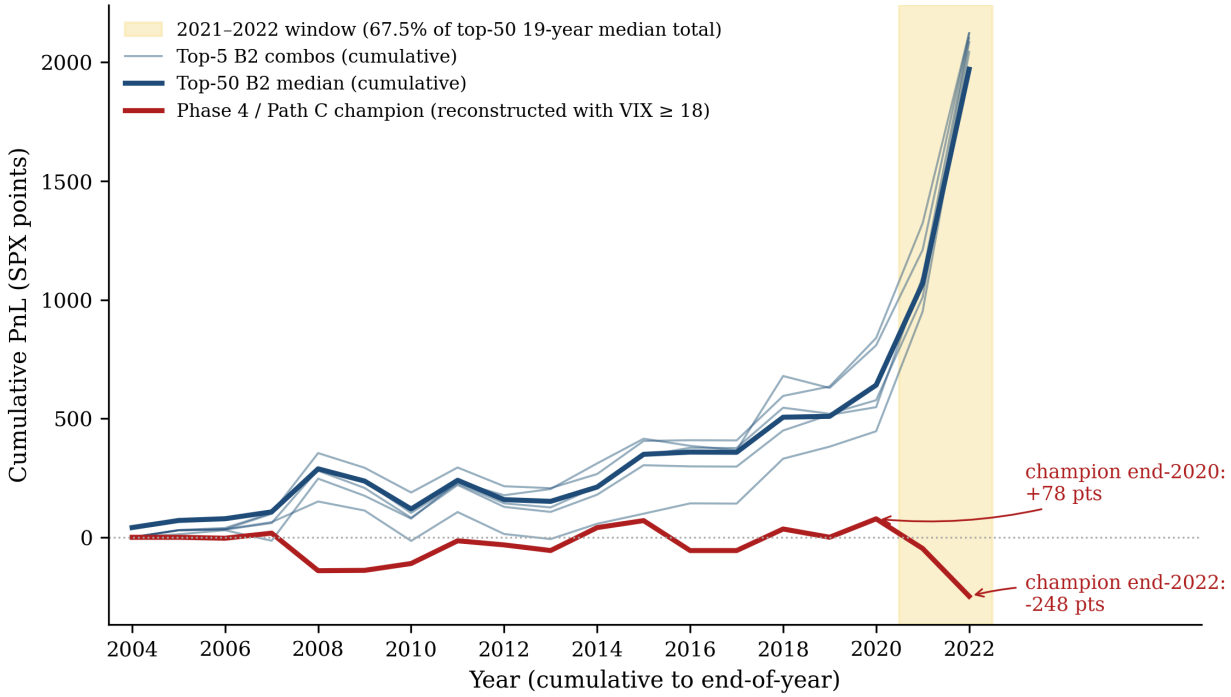
有四个反例值得明确点名。**2009** (VIX 约 31): top-50 100% 亏损, 中位 -52 点。**2010** (VIX 约 22): 100% 亏损, 中位 -118 点。**2012** (VIX 约 18): 100% 亏损, 中位 -82 点。**2017** (VIX 约 11): 100% 亏损, 中位 -1 点。这四年合起来覆盖了我们样本外 VIX 监测的全部范围,而在四年里每一个 top-50 配置都亏损。另一端, **2021** (VIX 约 19-20, 典型) 与 **2022** (VIX 约 26) 的中位分别为 +429 与 +901 ——但 2009 (VIX 约 31, 样本内最高年均) 是决定性的负。没有任何单调 VIX 规则能正确穿过这些数据点。这是 "regime 集中, 而非 VIX 集中" 论证的核心: 2021-2022 是一个具体的市场结构,而不仅是一段高 VIX 时期。

4.3.6 冠军重建

为了让冠军的 19 年轨迹直接可查,我们从 Task 2 的交易列表重建冠军: 在每笔交易级别上应用前一日 VIX ≥ 18 过滤器。Task 2 无过滤器的 2,237 笔中,有 951 笔通过过滤器,19 年合计 -257.0 点——与 Task B2 中 Mode C 格子报告的 -238.2 点接近 (小差距源于 Mode C 的 0.5 点滑点几何 vs Task 2 仅以点计的会计)。截至 **2020 年末**,冠军累计点 PnL 达到一个局部峰值 **+78 点** ——在 2020 年末实时运行该策略的研究者可以合理把它读作 "边际但为正"。接下来两年把 +78 变成 2022 年末的累计 **约 -248 点**: 冠军在 2021 与 2022 两年合计亏了约 325 点,而正是这两年 top-50 monoculture 产出了其

全部 19 年正合计。逐年分解明确: **冠军在 2021 亏 -125 点、2022 亏 -201 点** —— 这两年恰好是 *救活* top-50 monoculture 的两年。冠军并非只是在 2004-2022 上整体差;它是具体地在那个救活唯一胜出参数族的年份上差。

Figure 9. Cumulative PnL 2004–2022: top-50 combos concentrate gains in 2021–2022
Phase 4 champion ends the 19-year series at ≈−250 pts despite a positive 2007–2020 path

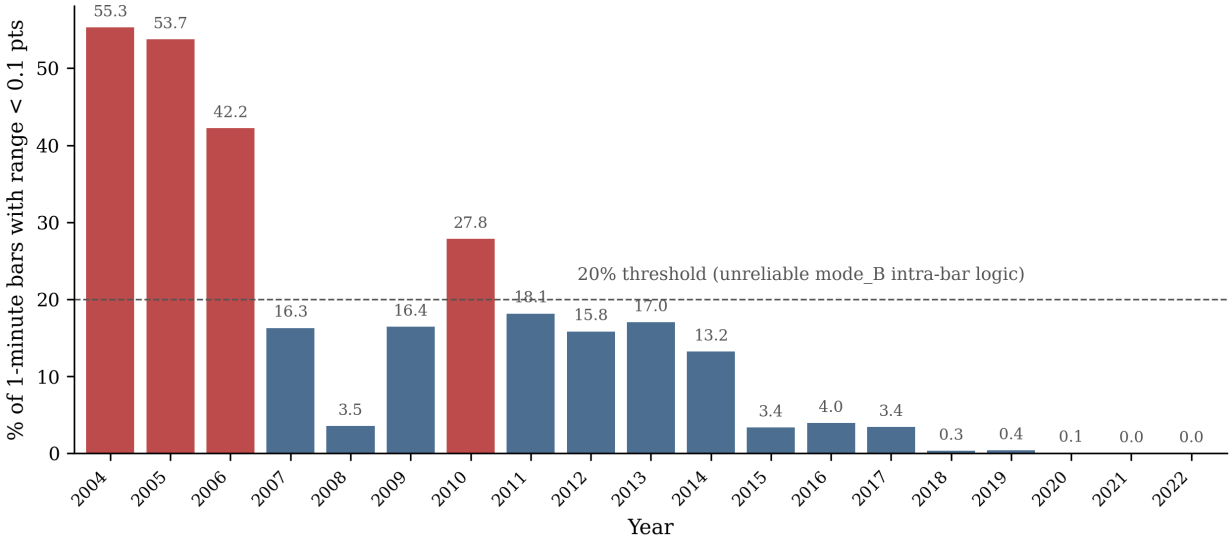


2004-2022 累计 PnL 曲线: B2 top-5 组合 (细灰线)、top-50 中位 (粗蓝)、重建的 Phase 4 冠军 (红)。top-50 中位从 2004 到 2020 基本持平偏降,然后在 2021-2022 急速上冲;Phase 4 冠军从 2004 到 2020 基本持平偏降,而在 2021-2022 *继续下降*。那段上冲贡献了 monoculture 19 年正合计的全部,而冠军并不参与。

图 9 把累计 PnL 曲线叠在一起。冠军曲线从 2004 到 2020 基本持平偏降, 2021-2022 继续下降。top-50 中位曲线从 2004 到 2020 也基本持平偏降,但在 2021-2022 急速上冲——那段上冲构成 monoculture 19 年正合计的全部,而冠军 *不* 参与那段上冲。

4.4 数据质量警示

Figure 8. SPX 1-minute bar degeneracy drops sharply after 2018; 2004–2006 exceed 40% 2021–2022 (the years driving top-50 PnL) have essentially zero degenerate bars



按年份 1 分钟 K 线范围 < 0.1 指数点 ("退化") 的百分比。20% 阈值 (虚线) 是我们采用的可靠性下限。2004 (55.3%)、2005 (53.7%)、2006 (42.2%) 远高于阈值;2007-2017 在 3.4% 到 27.8% 之间;2018 之后均低于 0.5%。2021 与 2022 (驱动 top-50 19 年正合计的两年) 的退化比例为 0.0%。2010 (4.0%) 相对其邻居 (2009: 12.2%; 2011: 27.8%) 偏低,反映该日历年入站 tick 节奏的短暂变化;由于 2010 已在 20% 可靠性阈值以下,我们将此作为观察而非需要修正的伪像。

图 8 显示按年份的 "退化" 1 分钟 K 线比例,定义为范围 < 0.1 指数点的 K 线。2004-2006 分别为 55.3%、53.7% 与 42.2% —— 都显著高于我们采用的 20% 可靠性阈值。那几年 Mode B 的秒内逻辑受到实质性影响,我们不在那里声明精度。但样本外结论 **不** 依赖 2004-2006: 剔除这三年后留下的 16 年样本 (2007-2022) 中,top-50 中位合计仍被 2021-2022 主导,四个全负年份 (2009、2010、2012、2017) 不受数据质量切分影响。

从 2018 年起, 退化 K 线比例都低于 0.5%; 驱动 monoculture 正合计的 2021 与 2022 两年, 退化比例为 0.0%。所以核心发现不是早期历史数据质量的产物。

一项相关验证 (Task A) 确认: **1 分钟** 模拟在 **rider** 家族上是 1 秒模拟的一个可靠代理,但在 **scalper** 家族上 **不是**。Mode C 在 1 分钟节奏下,对无过滤器 rider 合计相对 1 秒 ground truth 只偏 +0.6%;而 scalper 的 Mode C 相对 1 秒 ground truth 偏 -54.9%,因为紧凑的 `ts_act=3, ts=5, sl=5` 配置在 1 分钟节奏下无法被准确解析。这就是我们不在样本外报告 Phase 1/2 的原因: 仅 scalper 的 1 分钟与 1 秒差异本身就会淹没任何 regime 信号 (`reports/task_a_1m_validation.md`)。一个 scalper 样本外检验需要与 rider 样本外相同规模的 1 秒级处理,那是另外一个数量级的计算量,也是另一个独立工作流。

4.5 Path C 的重新解读

§3.4 的 MES STRONG_GO 作为一个正确的测量仍然成立: rider 确实在 2023-2026 上于每笔 MES 交易产生了净 $24.92, CI 不跨零。§4 否认的是对那个测量的 **解读**。rider 2023-2026 的 edge 是 2021-2022 regime 通过市场结构的持续性 (0DTE 流动性、dealer-gamma 定位、后 COVID 波动率)

被延伸到了前向。一旦样本外窗口把 regime 时钟重置回 2004,edge 就消失了。换言之,Path C Stage 2 每笔净 **+$24.92 MES** (等价于 **+$249.2 每手 ES**, 按 $/点 乘数 10:1 放大) 与其 CI 不跨零的判定,确认了一个配置的 *执行层可行性*,而这个配置的 *统计层 edge* 是一个两年的异常。

5. 讨论

5.1 我们做对了什么,做错了什么

在讨论失败之前,应该先精确说明哪些部分 **没有** 失败。模拟引擎兑现了它的正确性目标: Python 参考实现与 Rust 实现的对等测试套件端到端通过,1 秒因果时序全程得到执行,VIX 与 DR 过滤器只在已收盘 K 线信息上评估,每笔交易记录都带有独立的决策时间戳与执行时间戳。§3.4 的执行层分解——把 91.5% 的 SPXW-0DTE 摩擦归因于价差、只把 8.5% 归因于 theta ——就我们所见是一个对 "为什么即使底层点 PnL 为正, SPX 0DTE 期权仍是一个敌意场所" 的正确诊断。

失败的是研究协议本身。用我们自己的第一人称讲,三个具体错误是:

1. **我们把全部可用样本用在每一个决策上**。 3.1 年的 2023-2026 窗口同时被用在参数选择 (Phases 1-3)、过滤器选择 (Phase 4)、出场设计 (Phase 2 trailing、Phase 3 time-stop),以及 Path C 验证上。在写下 Path C STRONG_GO 的那一刻,没有任何干净的样本外保留。在 MES futures 上的 "验证" (§3.4) 是一个 **品种** 交叉检验,而不是 **样本** 交叉检验——底层的交易列表仍然是在同一批 780 天上选出来的。我们验证的是 "同一批交易在不同合约下 CI 为正",而不是 "不同样本确认这些交易来自一个正期望值分布"。
2. **我们把过滤器的 "稳健性" 当作跨组合稳定性来看待**。 Phase 4 报告 VIX ≥ 18 过滤器在 rider 家族上改善了盈亏比,我们把这读作过滤器捕捉到真实 regime 的证据。事实上,那个家族中的每一个组合都是在同一个样本上选出来的,所以跨组合稳定性只是联合选择,对事前样本外稳定性没有任何说明力。图 7 在样本外上 VIX-PnL 微弱的相关 ($R^2 = 0.083$),以及那四个覆盖 VIX 全范围的全负年份,才是我们从未要求过的证伪性证据。
3. **证据对我们有利时, 我们就停下了验证链路**。 Path C 给出了我们一路想要拿到的判定——一个在足够场所上可交易的 rider。我们当天就写下 STRONG_GO 文档。那时候 Task B2 的长历史网格搜索形式上被列为 "资本部署前的一项偏执检查";我们没有让 paper-trading 计划的任何部分以其结果为前提。样本外检验被当成了一个手续,而不是一个一票否决权。

这些不是冷僻的错误,是教科书式的后选择推断失败 (White 2000; Hansen 2005; Bailey, Borwein, Lopez de Prado, & Zhu 2014)。它们也是一个研究环境的副产品: 研究者控制管线的每一步,没有外部裁判;在收集 2004-2022 证据之前,我们从未把 2023-2026 的选择路径送给过任何外部方检视。Phase 4 冠军在样本外网格上排名 309,069/460,800,恰恰是一个在狭窄的 regime 特定样本上选出的 top-1 组合,在长样本多样化样本外中应当得到的结果。这不意外。这被流程预定。

5.2 Regime 集中 vs 纯过拟合

对我们的结果有两个竞争性诊断: **纯过拟合** (样本内参数被拟合到 2023-2026 具体的噪声实现,在任何更长样本上都不会推广) 与 **regime 集中** (rider 有一个在 2021-2022 的具体市场 regime 中才存在、在其它时期不存在的真实 edge)。证据偏向 regime 集中,而非纯过拟合:

- 2004-2022 top-50 是一个 *monoculture* (50/50 位于 tf=15, lb=20), 不是参数空间的随机散布。如果主导是过拟合,我们期望在不同样本上看不到稳定的顶部聚类。
- 2008 (+181)、2018 (+148)、2020 (+131)、2022 (+901) ——都与结构性波动率冲击相关——对 top-50 中位为正。背后机制 (dealer 再对冲事件中突破被延伸) 合乎情理。
- 2023-2026 每笔期望值 (+5.3 点, n=129) 落在 top-50 聚类 2021-2022 分布的内部, 不是异常值。
- 2009、2010、2012、2017 —— 高或中度 VIX 但无结构性波动冲击的年份 —— 对 top-50 全负,即便带过滤器也如此。过滤器没有识别出该 regime。

其含义是: 我们的冠军捕捉到的是一个真实但非平稳的现象。它不是 "一个策略",不是那种在合理 regime 上都保持正期望值的规则。它是一个 *事件驱动* 效应,最多只能作为一个在识别出结构性波动事件期间的主观叠加层。

5.3 缺少样本外保留样本的代价

如果我们在 Phase 1 开始之前就把 2004-2022 样本中的五年 (例如 2018-2022) 保留为一个未触及的 holdout,那么 Phase 4 冠军重建在 2021+2022 的 -257 点在我们跑 Path C 之前就会可见,整个 Path C 工作流也会被一票否决。所以缺少样本外保留的代价就是 Phase 2-5 加上 Path C Stage 1-2 的全部成本 —— 大约六个研究员工时 (researcher-weeks) 加上一次 $5.50 的 Databento 拉取 —— 换来 "已经测过负结果" 的确定性。Databento ES 长历史拉取 (最初作为 Task 4, $20 预算) 在 2026-04-22 收到 Task B2 结果后被预先取消: 机制层面的问题已在 SPX 1 秒保真度上被解决,更高保真的确认只会在更高代价上重新测出同一个负结果。

在我们这个具体案例中,一个修正后的协议会是:

1. 把 2004-2022 分成一个 **训练** 窗口 (2004-2013 含) 与一个 **保留样本** (2014-2022)。保持保留样本未触及。
2. 仅在训练窗口上运行 Phase 1-5 与 Path C。明确禁止查看保留样本的指标。
3. 在 Phase 4 / Path C 结束时,对保留样本应用一个 **预登记的接受标准**。例如: "冠军必须在 ≥ 7/9 保留样本年份上产生正年度 PnL, 其中 ≥ 5/9 年份的逐年 CI 不跨零, 且保留样本平均 Sharpe 在 95% 水平上不跨零。" 不通过 → 策略被拒;通过 → 进入实盘部署。
4. 实盘部署之后, 按月把正向 PnL 与预登记的保留样本 PnL 分布做对比。若正向 PnL 显著低于保留样本分布, 触发自动复盘。

这是一个常规的滚动前推 + 隔离期保留样本 (walk-forward + embargoed holdout) 设计 (López de Prado 2018, 第 7 章)。其中没有任何新东西;我们这种情形的独特在于: 纪律没有被外部强加, 我们自己也没有强加给自己, 于是它就根本没有被强加。

5.4 为什么 Path C 没有发现问题

对 §5.1 叙事的一个自然反驳是: "Path C 在不同品种上做过验证——那不就是一个样本外检验吗?" 不是,这个区分值得仔细说明。

跨品种验证 (cross-instrument validation) 检查的是: 同一份从 SPX 指数映射到可交易合约 (MES 或 ES) 的交易列表,在新的摩擦面下是否仍保留 edge。它是回答 "我们能以可承受成本交易它吗?" 的正确检验。它是回答 "底层信号是真实的吗?" 的 *错误* 检验。信号有效性问题要求的是一份 **新的** 交易列表,来自信号从未暴露过的数据。

我们的 Path C 交易列表是 2023-2026 样本内过滤器选出的同一批 129 笔交易;MES 只是在新的成本模型下对这 129 笔重新定价。这种重新定价有三个有用性质 (摩擦小、与 SPX 相关性高、edge 捕获率约 99%),但它们都没有说明信号的样本外稳定性。若样本内那 129 笔交易来自一个 regime 绑定分布,那么 MES 重新定价后的 129 笔仍然来自该分布,因为交易本身就是同一批。这在结构上类似于把 "我的回测无 bug" (Path C 回答的问题) 误当作 "我的回测可推广" (只有新样本才能回答的问题)。

我们现在把 Path C 看作验证的 *必要但不充分* 部分: 它排除了由摩擦引起的失败, 但不能确立信号有效性。把 Path C 完全从项目中剥掉并不能修正主要错误;加入一个预登记的样本外保留样本才能。

5.5 局限性

- **模拟保真度**。 Mode C 的 0.5 点滑点是一个简单模型。对 rider 持仓 (20-45 分钟) 而言影响小; 对 Phase 1/2 scalper 家族, 一项验证研究 (`reports/task_a_1m_validation.md`) 显示 1 分钟样本外模拟对 scalper 偏 ±15% ——这是我们不在样本外报告 Phase 1/2 的一个原因。
- **单一品种样本外**。 样本外检验仅在 SPX 指数点上进行。我们没有在 MES futures 上做样本外 (Databento 成本增加 $20), 理由是 SPX 1 秒是信号的标准源, §3.4 的 MES 翻译建立的是 *摩擦* 可行性, 而非 *信号* 有效性。
- **无正式 reality check 检验**。 我们报告了单一配置在 460,800 格网格中的排名, 以及 top-50 选择在特定年份的集中性, 但没有应用正式的 Superior Predictive Ability 检验 (Hansen 2005) 或 deflated Sharpe ratio (Bailey & Lopez de Prado 2014)。鉴于差距的量级 (冠军排在自顶 67 分位, 网格中位 -133 点), 我们不认为正式检验会改变判定, 但它会让证伪在定量上更精确。
- **Scalper 家族未在样本外检验**。 Phase 1-2 scalper 冠军在 2023-2026 上点 PnL 为正 (+5,732 点), 被 SPXW 价差消耗。MES 摩擦更小, 但 scalper 从未在 MES 上被测试过 (也未在 2004-2022 以 1 秒保真度测试过)。scalper 样本外检验是独立工作流;Task A 显示 1 分钟样本外不是有效的捷径。Task G 综合把它列为延期项而非可能改变主判定的开放问题。
- **无显式因果过滤器稳定性检验**。 我们从四个全负年份与微弱 VIX 回归论证过滤器没有识别出稳定 regime, 但没有预登记一个具体的过滤器稳健度指标。更严格的工作流会是, 例如 "要求 top-N 组合在 ≥ 15/19 年上为正且中位 > 0" 作为样本外结果的接受标准, 应用于任何进一步部署决策之前。
- **Regime 识别本身是事后的**。 "2021-2022 dealer-gamma / 0DTE 流动性" 叙事是从逐年分解反推出来的,不是预登记的。一个有纪律的前向检验应当在断言 "策略在 regime X 内有效" 之前就 (以事前可观察方式) 定义 regime 边界——否则 "edge 是真实的, 它只在 regime X 中出现" 本身就是参数选择之上的第二轮选择,而参数选择正是本文所批评的。我们在此标记这一点,

以劝阻读者把 §5.2 的 regime 集中诊断当作 "出现类似 2021-2022 条件时就重新部署冠军" 的通行证。

6. 结论

我们报告一个负结果与一个方法论教训。一个在 2023-2026 数据上选出的 SPX Donchian 突破 rider——带前一日 VIX 过滤器与保守 trailing 出场, 并在 MES futures 上确认了品种可行性——无法在 19 年样本外检验中幸存。其 2023-2026 的 edge 是一个 regime 集中于 2021-2022 的异常的前向延伸;参数本身在样本外样本上排名自顶 67 分位, 且当过滤器生效时, 在 2021-2022 两年亏损。

我们给出以下指引作为实践者的参考,每一条都与我们自己的经验一致,但都未在多个独立策略项目上预登记验证过:

1. **在任何选择步骤之前,先预登记样本外窗口**。 对于在日度或日内数据上运行、regime 周转尺度在宏观周期量级上的策略,五年可以作为一个可辩护的最低值;对于机制可能依赖具体市场结构 regime 的策略,更长更安全。
2. **在保留样本上验证 edge,不要在一个源自同一样本内交易列表的不同品种上验证**。 执行层交叉检验 (SPX → SPXW → MES) 验证的是摩擦,不是 edge。两者都必要;只有其中一项是目前大多数研究协议所具备的。
3. **把资本部署条件化在样本外检验上, 而不是执行层翻译上**。 我们写下的 Path C STRONG_GO 判定在技术上正确 (MES 上摩擦可控),在策略上却是误导性的 (edge 是 regime 绑定的)。这句话的两部分同样重要。
4. **事先考虑一个正稳定性的接受标准**。 一个可行的默认 (作为起点而非普适规则) 是要求排名靠前的配置在至少 75-80% 的样本外年份上产生正年度 PnL, 且 CI 不跨零。这会过滤掉那些被少数年份救活的 monoculture。具体阈值应取决于策略的预期持仓期与合理的 regime 周转频率。

对本文所述的研究项目,工作流已关闭。下一个策略不会再把 2004-2022 SPX 1 秒样本用于其选择循环中——那个样本现在是我们唯一还未触及、可以用于一次实际样本外检验的资源。该配置不曾被部署,也不会被部署。

正面的解读是: 协议起作用了。一个 3.1 年的样本内选择弧线产出了一个候选;一系列执行层验证精炼了品种选择;一个长历史样本外检验在资本部署之前就否决了判定。总成本: 约六个研究员工时、约 $5.50 Databento 拉取、加上样本外网格的 8.1 小时计算。替代方案——给一个 regime 集中的配置启动 paper-trading, 在下一次市场结构切换时发现其失败——要贵上几个数量级。在这个意义上,负结果正是一套严谨研究流程应当产出的结果,而公开地记录这个结果,是本文试图完成的工作。

Acknowledgments

Interactive Brokers, Cboe, and Databento for data. The simulation engine was built in Python with NumPy, pandas, and PyArrow; Rust parity uses PyO3 and maturin; grids were parallelized with `concurrent.futures.ProcessPoolExecutor`.

References

Bailey, D. H., Borwein, J., Lopez de Prado, M., & Zhu, Q. (2014). "Pseudo-mathematics and financial charlatanism: the effects of backtest overfitting on out-of-sample performance." *Notices of the American Mathematical Society*, 61(5), 458-471.

Bailey, D. H., & Lopez de Prado, M. (2014). "The deflated Sharpe ratio: correcting for selection bias, backtest overfitting, and non-normality." *Journal of Portfolio Management*, 40(5), 94-107.

Bailey, D. H., Borwein, J., Lopez de Prado, M., & Zhu, Q. (2016). "The probability of backtest overfitting." *Journal of Computational Finance*, 20(4), 39-70.

Hansen, P. R. (2005). "A test for superior predictive ability." *Journal of Business & Economic Statistics*, 23(4), 365-380.

Harvey, C. R., Liu, Y., & Zhu, H. (2016). "... and the cross-section of expected returns." *Review of Financial Studies*, 29(1), 5-68.

Leinweber, D. J. (2007). "Stupid data miner tricks: overfitting the S&P 500." *Journal of Investing*, 16(1), 15-22.

Lempérière, Y., Deremble, C., Seager, P., Potters, M., & Bouchaud, J.-P. (2014). "Two centuries of trend following." *Journal of Investment Strategies*, 3(3), 41-61.

Lo, A. W., & MacKinlay, A. C. (1990). "Data-snooping biases in tests of financial asset pricing models." *Review of Financial Studies*, 3(3), 431-467.

López de Prado, M. (2018). *Advances in Financial Machine Learning*. Wiley.

Sullivan, R., Timmermann, A., & White, H. (1999). "Data-snooping, technical trading rule performance, and the bootstrap." *Journal of Finance*, 54(5), 1647-1691.

White, H. (2000). "A reality check for data snooping." *Econometrica*, 68(5), 1097-1126.

数据产物索引

所有定量主张都引用 `results/` 或 `data/` 下的具体产物。主要产物如下:

主张 (§)	产物
Phase 1 冠军 (§3.1, 表 3)	`results/run_20260420_162326_full_grid_phase1/metrics.parquet`
Phase 2 冠军 (§3.2)	`results/run_20260420_235118_full_grid_phase2/metrics.parquet`
Phase 4 冠军 (§3.3, 表 3)	`results/run_20260421_033134_full_grid_phase4/metrics.parquet`
MES Path C Stage 2 (§3.4, 图 3, 表 3)	`results/path_c_stage2/summary_2_4.json`
摩擦分解 (§3.4, 图 2)	`scripts/options_validation_stage2_cost.py` 输出
VIX30 vs VIX1D (§4.1, 图 4)	`data/vix_data/`, `data/vix1d_data/VIX1D_1min_full_history.parquet`
460,800 组合网格 (§4.3, 图 5, 表 2)	`results/task_b2_rider_grid/metrics.parquet`
Top-50 monoculture (§4.3.2, 表 1)	`results/task_b2_rider_grid/report_tables/top_mode_c_by_total_pnl.parquet`
冠军邻域 (§4.3.3, 图 6)	`results/task_b2_rider_grid/metrics.parquet`
逐年 top-50 (§4.3.4, 图 9)	`results/task_b2_rider_grid/report_tables/per_year_pnl_top.parquet`
VIX 回归 (§4.3.5, 图 7)	`results/task_b2_rider_grid/report_tables/top50_year_loss_summary.parquet`
K 线质量 (§4.4, 图 8)	`results/task_b2_rider_grid/bar_quality.json`
冠军逐年重建 (§4.3.6, 图 9)	Task 2 trades + VIX history; 代码在 `reports/paper/generate_figures.py`
Task G 综合 (§4.5, §5)	`reports/task_g_synthesis.md`

模拟源代码: `src/spx_donchian_hf/` (Python 参考) 与 `rust_engine/` (Rust 对等)。图表生成: `reports/paper/generate_figures.py` 与 `reports/paper/generate_tables.py`。

---

1. 以 SPX 指数点计, Mode C 的 "每笔 PnL" **已经扣除了 1.0 点的往返滑点** ——在报告这些每笔数字 (例如 Phase 4 +5.33 点/笔) 时不再另外扣除。在 Path C 的 MES 表格中,每笔美元数字还另外扣除了 $4 的往返摩擦模型 (单边 $2 MES 手续费 + 价差代理),这一点在相关轴标签上显式标明。 ↩
2. 全文中, "Phase 1 冠军" 指 `tf=1, lb=5, sl=5, tp=10, ts=off, time_stop=25`; "Phase 2 冠军" 指 `tf=1, lb=5, sl=5, tp=40, ts_act=3, ts=5, time_stop=25`; "Phase 4 冠军" 指 `tf=30, lb=50, sl=off, tp=off, ts_act=10, ts=10, time_stop=off, vix=18, dr=off`。除非另有说明,三者均使用 Mode C 成交。 ↩

3. 表 1 的列是 2004-2022 每个日历年加上 `total`。(a) 10 行全部共享 `tf=15, lb=20` —— §4.3.2 总结过的 monoculture 特征。(b) PnL 对比主要出现在 2021-2022 (绿色, 大正) 与 2009/2010/2012/2017 四个全负年份 (红色, 全部为负);若用颜色梯度渲染,得到的视觉效果与正文的逐年故事吻合,不会提供独立证据。 ↩