Trading Research Compendium

ES Futures / SPX Options / Volatility

March 2026

Compiled by Luther for Daniel

10 Research Papers + Backtests + Signal Ideas

VIX Index Decomposition (Cboe, Ed Tom)

Source

* Author: Edward K. Tom, Sr. Director, Derivatives Market Intelligence, Cboe
* Date: August 1, 2025
* Type: Official Cboe whitepaper (not academic)

Core Concept

Decomposes VIX moves into 6 principal components that explain WHY VIX moved, not just how much:

The 6 Factors

1. Expected Move per Sticky Strike -- the VIX change already priced into the skew given an SPX move. Dominant factor in normal regimes (VIX <20, moves <1pt = 88% of VIX move explained)
2. Parallel Shift -- wholesale repricing of optionality across ALL strikes. Occurs in 2+ sigma shocks. Directly proportional to uncertainty of new catalyst. Key driver of "VIX overreaction"
3. Put Skew Gradient (15-45 delta puts) -- marginal demand for liquid OTM protective puts ("shoulders" of skew)
4. Call Skew Gradient (15-45 delta calls) -- marginal demand for liquid OTM calls
5. Downside Convexity (1-15 delta puts) -- demand for tail risk/crash protection ("wings")
6. Upside Convexity (1-15 delta calls) -- demand for levered upside/rally options ("wings")

Key Methodology

* Interpolates front/back month SPX skews into single 30-day fixed-strike skew (variance space interpolation)
* Perturbation analysis: each factor's contribution = marginal VIX change from that factor alone
* Factors are additive: sum of 6 components ~= total VIX change

Critical Findings

Regime Behavior

* VIX <20, move <1pt: Sticky strike explains 88% of moves. Other factors negligible.
* VIX >25, move >5pts: Parallel shift becomes dominant. This is when "VIX is broken" narrative appears.
* Positive co-movement (SPX up + VIX up): Occurs 20% of the time. Almost always driven by parallel shift.

Case Study: Yen-Carry Unwind (Aug 5, 2024) vs Liberation Day (Apr 4, 2025)

Both: VIX doubled from ~23 to ~40-45 on 3-6% SPX decline. BUT decomposition reveals completely different sentiment:

Yen-Carry (bearish signal):

* Small sticky strike (low vol premium priced in)
* HUGE parallel upshift (uncertainty repricing)
* Steep put skew + downside convexity bid (crash protection buying)
* Negative call skew + convexity (selling calls to fund puts)

Liberation Day (bullish signal):

* Large sticky strike (higher vol premium already priced in)
* Large parallel upshift (uncertainty)
* BUT: massive bid for upside convexity -- traders buying levered calls, positioned for rally
* Market rallied +7% in 2 days, +14% in 30 days

The Actionable Insight

Same VIX spike, opposite market implications. The decomposition revealed Liberation Day traders were betting on recovery, not

hedging for crash. This is a tradeable signal.

Applicability to Our System

HIGH PRIORITY -- Intraday Trading Signal

1. We have live OPRA SPX/SPXW data -- can compute this decomposition in real-time
2. VIX spike classification: When VIX spikes, decompose WHY to determine if it's crash-hedging (bearish) or opportunity-buying (bullish)
3. Put/call convexity ratio as sentiment: Ratio of downside to upside convexity demand = directional signal
4. Parallel shift as regime change detector: Large parallel upshifts = new uncertainty = widen stops

Implementation Approach

* Build interpolated 30-day SPX skew from live OPRA data (front + back month)
* Track each component intraday: sticky strike contribution, parallel shift, put/call gradient, convexity
* Generate alerts when decomposition diverges from "expected" pattern
* Key signal: upside convexity bid during VIX spike = bullish contrarian

Data Requirements

* SPX/SPXW option chain with IVs across full strike range ? (have via OPRA live stream)
* SPX spot price ? (have via Yahoo)
* Delta mapping for strike classification (need to compute from IV + time)
* Historical skew snapshots for comparison (need to build)

Relation to Existing Signals

* VIX Term Structure signal: This decomposition adds granularity -- WHAT's driving VIX, not just the level
* GEX signal: Complementary -- GEX measures dealer hedging exposure, this measures investor sentiment
* Put/Call OI ratio: This is a more sophisticated version -- decomposed by delta bucket

Notes

* Cboe plans to launch a web tool for daily decomposition (Fall 2025)
* Historical decompositions for top 25-year VIX moves to be published
* Contact: etom@cboe.com, mxu@cboe.com

Overreaction / Momentum (arXiv 2602.18912)

Date: 2026-03-02
Asset: ES futures (E-mini S&P 500), 1-min bars
Data: Jan 2, 2025 -> Feb 19, 2026 (292 RTH days, 111K bars)
IS/OOS Split: Oct 17, 2025 (204 days IS / 88 days OOS)

Paper Summary

The paper proposes that vol-normalized short-term returns ("overreactions") predict subsequent price movement. In equities (AAPL), overreactions at 5-min scale show momentum continuation. The question: does this transfer to ES futures?

Methodology

* Signal: Z-score on 5-min returns, normalized by rolling 1hr (60-bar) volatility of 1-min returns
* Overreaction threshold: |z| > 2.0 (also swept 1.5-4.0)
* Strategy: Fade (reversal) -- short after UP overreaction, long after DOWN overreaction
* Hold period: 5 bars (5 min), also swept 1-25 bars
* Transaction costs: 0.7 bps RT (0.25pt spread + $1.24/side commission)
* Validation: IS/OOS split, non-overlapping trades, daily PnL Sharpe

Key Finding: ES Overreactions Are Weak, Not Strong Reversals

Previous run claimed: 91% reversal rate after UP overreaction, -71 bps, Sharpe ~15.9
Actual finding: Those numbers were severely inflated. The true picture:

Continuation Analysis (OOS, z > 2.0)

Direction	Horizon	Reversal %	Avg Fwd (bps)	t-stat	Verdict
UP ?	5-min	51.2%	-0.4	-1.27	NEUTRAL
UP ?	15-min	50.5%	+0.6	+1.00	NEUTRAL
UP ?	25-min	49.2%	+1.7	+2.20	NEUTRAL (slight continuation!)
DOWN ?	5-min	52.6%	+0.8	+2.33	SLIGHT REVERSAL
DOWN ?	15-min	54.0%	+0.7	+1.18	NEUTRAL
DOWN ?	25-min	55.6%	+1.4	+1.68	SLIGHT REVERSAL

The "91% reversal" was wrong. Actual reversal rates are 50-56%, barely above coin flip.

Why the Previous Sharpe (~15.9) Was Inflated

1. Signal clustering: 49-54% of signals occur within 5 bars of each other. Counting each as an independent trade massively overstates degrees of freedom
2. Per-trade annualization: Using trades_per_year x mean/std inflates Sharpe when trades cluster
3. Proper metric is Daily PnL Sharpe -- aggregates all trades within a day, then annualizes

Reversal Strategy OOS Results (z > 2.0, hold 5 bars)

Strategy	N Trades	Win Rate	Avg Trade	Daily Sharpe	Profit Factor
Fade UP (short)	484	49.8%	-0.27 bps	**-0.91**	0.90
Fade DOWN (long)	627	50.6%	+0.10 bps	**+0.33**	1.03

Neither is tradeable. Fade UP loses money. Fade DOWN is marginally positive but with Sharpe 0.33.

Non-Overlapping Trades (OOS, removes clustered signals)

Strategy	N Trades	Win Rate	Avg Trade	Daily Sharpe	Profit Factor
Fade UP	253	49.0%	+0.04 bps	**+0.12**	1.01
Fade DOWN	315	53.7%	+0.37 bps	**+1.12**	1.12

Fade DOWN non-overlapping shows Daily Sharpe 1.12 -- this is the most interesting result, but:

* Only 88 OOS days -> wide confidence interval
* Average trade is +0.37 bps ~= $1.11/trade at 1 lot -- barely above costs
* Profit factor 1.12 is thin -- one bad week erases a month of gains

Parameter Sweep Highlights (OOS)

Best OOS configs are at extreme thresholds (z > 3.5) with tiny sample sizes (n=17). Not reliable.

The only config with decent sample AND positive performance:
z=2.0, hold=25 bars, FADE DOWN: Sharpe 0.57, WR 54.3%, avg +0.65 bps, n=554

But this doesn't replicate in IS (IS Sharpe was -0.01 for same config).

IS -> OOS Stability

Config	IS Sharpe	OOS Sharpe	Verdict
z=2.0 h=5 Fade DOWN	-0.66	+0.33	? Sign flip
z=2.5 h=15 Fade DOWN	+0.67	-1.30	? Sign flip
z=2.0 h=25 Fade DOWN	-0.01	+0.57	? IS was flat
z=1.5 h=10 Fade DOWN	-0.17	-1.41	? Both negative

No configuration shows consistent IS -> OOS positive performance. The signal is not robust.

Asymmetry: DOWN > UP

One genuine finding: DOWN overreactions show slightly more reversal tendency than UP across almost all horizons and thresholds. This could reflect:

* Dip buying in a structurally bullish market (2025 was up ~20%)
* Passive rebalancers providing mean-reversion flow after selloffs
* Short covering amplifying bounces

But the edge is 0.1-0.4 bps per trade -- below practical trading threshold.

Verdict

? NOT TRADEABLE STANDALONE

The overreaction signal in ES is:
1. Real but tiny -- effect sizes of 0-1 bps, well below transaction costs
2. Unstable -- no IS/OOS consistency at any parameter combination
3. Asymmetric -- DOWN overreactions reverse slightly more than UP, but not enough
4. Clustering-prone -- signals bunch together, inflating apparent sample sizes

Potential Use as Aggregator Input

The signal could contribute as a weak feature in a multi-signal aggregator:

* DOWN overreaction z-score as a "dip-buy tendency" indicator (weight: low)
* Combined with volume surge, orderflow, or cross-asset confirmation
* NOT worth a standalone signal script -- too weak to justify complexity

Why It Works in AAPL (Paper) But Not ES

* AAPL has wider spreads (relative) -> larger overreaction amplitude
* Single-stock microstructure is noisier -> bigger vol-normalized moves
* ES is the most efficient market in the world -- arb squeezes anomalies to near-zero
* ES mean-reversion at 5-min scale is already well-arbitraged by HFT

Signal Script Decision

? No signal script created. The signal does not meet the minimum bar:

* Daily Sharpe < 1.0 consistently
* Average trade < 1 bps after costs
* No IS/OOS stability

If future work combines this with orderflow or sentiment data and finds amplification, revisit.

Files

* Backtest script: `scripts/overreaction_paper_backtest.py`
* Previous iterations: `scripts/backtest_overreaction_v{1-4}.py`, `scripts/es_overreaction_backtest.py`
* Data: `data/es_1min_bars.csv`

FactorMiner: Regime-Switching Factors (arXiv 2602.14670)

Paper: arxiv 2602.14670 | Score: 43 | Date: 2026-03-02

TL;DR for ES Intraday Trading

FactorMiner provides 110 formulaic alpha factors validated on 10-min intraday bars. The regime-switching factors (using IfElse with volatility/skewness/kurtosis conditions) are the strongest performers, with the highest XGBoost feature importance and stepwise selection priority. These are directly adaptable to ES futures intraday trading.

---

1. What We Already Had

* Basic paper metadata in research-papers.json
* Knowledge that Appendix P has 110 factor formulas
* Understanding that regime-switching factors were strongest

2. What's New in This Extraction

A. Complete Operator Library (60+ operators, Table 2)

```
| Category | Key Operators | ES Relevance |
|----------|--------------|--------------|
| Arithmetic | Add, Sub, Mul, Div, Neg, Log, SignedPower | Basic building blocks |
| Statistical | Mean, Std, Skew, Kurt, Med | **Critical** -- Skew/Kurt drive regime detection |
| Time-series | Delta, TsRank, TsMax, TsMin, Delay, TsDecay | Core temporal patterns |
| Cross-sectional | CsRank, Scale | Less relevant for single-instrument ES |
| Smoothing | SMA, EMA, WMA | Standard |
| Regression | Slope, Rsquare, Resi | **High value** -- trend reliability signals |
| Logical | IfElse, Greater, Less, And, Or | **Key innovation** -- regime switching |
```

B. Regime-Switching Methodology (The Core Innovation)

The paper's key insight: use higher-order moments (Skew, Kurt) and volatility state as conditional switches to adapt factor logic to current market regime.

Framework:
```
IfElse(REGIME_CONDITION, FACTOR_A, FACTOR_B)
```

Where REGIME_CONDITION is typically:
1. `Greater(Std($returns, 12), Mean(Std($returns, 12), 48))` -- High volatility regime
2. `Greater(Abs(Skew($returns, 24)), threshold)` -- Extreme skewness regime
3. `Greater(Kurt($returns, 24), 3.0)` -- Fat-tail (leptokurtic) regime
4. `And(vol_condition, kurt_condition)` -- Compound regime detection
5. `Or(vol_extreme, price_extreme)` -- Any-extreme regime

C. Top Regime-Switching Factors (Ranked by XGBoost Importance + IC)

Factor 046 -- Volatility-Regime Reversal Divergence (IC=0.109, ICIR=0.94)
```
IfElse(
  Greater(Std($returns, 12), Mean(Std($returns, 12), 48)),
  Neg(CsRank(Delta($close, 3))),                    # High vol -> price reversal
  Neg(CsRank(Div(Sub($close,$low), Add(Sub($high,$low), 0.0001))))  # Low vol -> range position
```

```
)
```

ES Translation: When 12-bar volatility exceeds its 48-bar average -> short-term price reversal. When vol is normal -> mean-revert from range position.

Factor 042 -- Regime-Switching Skew Factor

```
IfElse(
  Greater(Abs(Skew($returns, 24)), 1.0),
  Neg(CsRank($returns)),           # Extreme skew -> reversal
  Neg(CsRank(Skew($returns, 24)))  # Normal -> skew-based signal
)
```

ES Translation: When return distribution is highly skewed -> fade the recent move. Otherwise -> position against the skew direction.

Factor 095 -- Higher Moment Regime Switch (IC=0.062)

```
IfElse(
  Or(Greater(Abs(Skew($returns, 24)), 1.5), Greater(Kurt($returns, 24), 4.0)),
  Neg(Resi($close, 6)),            # Extreme moments -> residual reversal
  Neg(TsRank($returns, 24))        # Normal -> momentum reversal
)
```

ES Translation: When skew > 1.5 OR kurtosis > 4 -> trade the regression residual. Normal conditions -> momentum-based signal.

Factor 094 -- And HighVol LowKurt Switch

```
IfElse(
  And(Greater(Std($returns, 12), Mean(Std($returns, 12), 60)), Less(Kurt($returns, 24), 2)),
  Neg(Delta($close, 3)),           # High vol + platykurtic -> price reversal
  Neg(TsRank($returns, 24))        # Otherwise -> rank momentum reversal
)
```

ES Translation: High volatility + thin tails (trending, not spiking) -> short-term reversal. Otherwise -> longer-term momentum reversal.

Factor 080 -- Rsquare Resi Adaptive

```
IfElse(
  Greater(Rsquare($close, 24), 0.7),
  Neg(CsRank(Slope($close, 24))),  # Strong trend (R^2>0.7) -> slope reversal
  Neg(CsRank(Resi($close, 12)))    # Weak trend -> residual reversal
)
```

ES Translation: Strong linear trend -> fade the slope. No trend -> fade the deviations from regression.

Factor 079 -- Regime Vol Range Pos Switch (Stepwise rank #3)

```
IfElse(
  Greater(Std($returns, 12), EMA(Std($returns, 12), 48)),
  Neg(Div(Sub($close,$low), Add(Sub($high,$low), 1e-6))),  # High vol -> short near high of range
  Div(Sub($high,$close), Add(Sub($high,$low), 1e-6))       # Low vol -> buy near high of range
)
```

ES Translation: Volatility expansion -> mean-revert from range highs. Low vol -> breakout/continuation from range highs.

D. Experience Memory -- Recommended Patterns (Table 4)

Pattern	Description	Success Rate
Higher Moment Regimes	Skew/Kurt as IfElse conditions for reversal	**High**
PV Corr Interaction	Price-volume correlation + amount efficiency	**High**
Trend Regression Adaptive	Rsquare/Slope/Resi operators; High R^2->slope reversal, Low R^2->residual reversal	**High**
Logical 'Or' Extreme Regimes	Or operator integrating multiple extreme indicators	**High**
Kurtosis Regime	Kurt to identify fat-tail environments, adjust reversal windows	**High**
Amt Efficiency Rank Interaction	Amount efficiency time-series rank with kurtosis	Medium

E. Forbidden Directions (Avoid These)

Direction	Problem
VWAP Deviation variants	Overcrowded signal space (40+ rejected candidates)
Standardized Returns	High correlation with existing factors
Simple Delta Reversal	Too simple, already captured
WMA/EMA Smoothed Efficiency	Correlation > 0.9 with existing

F. Admission Criteria (Section B)

* IC threshold: ?_IC >= 0.04
* Correlation threshold: ? < 0.5 (inter-factor)
* Replacement: IC >= 0.10 AND IC >= 1.3x existing factor AND only 1 correlated factor

G. Key Performance Numbers

* Best single factor IC: 0.109 (Factor 046, regime-switching)
* IC-weighted combination (110 factors): IC=0.150, ICIR=1.24
* XGBoost selection (all 110): IC=0.163, ICIR=1.49, Win Rate=92.6%
* Lasso needs only 8 factors to capture 95% of IC improvement
* Perfect monotonicity (1.0) across all combination methods

---

3. ES Intraday Adaptation Strategy

Single-Instrument Adaptations Required

The paper uses cross-sectional ranking (CsRank) across hundreds of stocks. For single-instrument ES:

* Replace CsRank with time-series z-score or percentile rank over rolling window
* Replace $amt with $volume (ES doesn't have separate amount field)
* Replace $vwap with session VWAP or rolling VWAP calculation

Most Actionable Factors for ES (Priority Order)

Tier 1 -- Directly Implementable:

1. Volatility regime switch (Factor 046 pattern): `if std(returns, 12bars) > mean(std, 48bars) then reversal else range_position`
2. Skewness regime (Factor 042 pattern): `if |skew(returns, 24bars)| > 1.0 then reversal else skew_signal`
3. R^2 adaptive (Factor 080 pattern): `if rsquare(close, 24bars) > 0.7 then fade_slope else fade_residual`
4. Kurtosis regime (Factor 045/087 pattern): `if kurt(returns, 24bars) > 3.0 then candle_reversal else range_reversal`

Tier 2 -- Requires Adaptation:

5. Higher moment compound (Factor 095): Uses Or(skew>1.5, kurt>4.0) as compound regime
6. Vol+Kurt compound (Factor 094): And(high_vol, low_kurt) detects trending (non-spiking) vol expansion
7. Trend reliability (Factor 083): And(R^2>0.5, low_vol) -> trend follow; else -> reversal
8. Candlestick regime (Factor 036): Vol regime determines whether to use shadow ratio or returns

Tier 3 -- Novel Concepts:
9. Resi acceleration (Factor 100): `Neg(Resi(Delta(Resi($close, 24), 3), 12))` -- second-order detrending
10. Vol-of-volume (Factor 077): Volatility of relative volume as reversal weight

Window Size Translation (10-min bars -> ES bars)

The paper uses 10-min bars with windows like 12, 24, 48. For ES:
* 1-min bars: multiply by 10 (so 12 -> 120, 24 -> 240 bars)
* 5-min bars: multiply by 2 (so 12 -> 24, 24 -> 48 bars)
* 10-min bars: use directly
* Hourly bars: divide by 6 (so 24 -> 4, 48 -> 8 bars) -- likely too coarse

Recommended ES Implementation (1-min bars, RTH only = 390 bars/day)

```python
vol_12 = rolling_std(returns, 120)        # ~2 hours
vol_48_avg = rolling_mean(vol_12, 480)    # ~8 hours (2 days)
skew_24 = rolling_skew(returns, 240)      # ~4 hours
kurt_24 = rolling_kurt(returns, 240)      # ~4 hours
rsq_24 = rolling_rsquare(close, 240)      # ~4 hours

high_vol = vol_12 > vol_48_avg
extreme_skew = abs(skew_24) > 1.0
fat_tails = kurt_24 > 3.0
strong_trend = rsq_24 > 0.7

if high_vol:
    signal = -delta(close, 30)            # Short-term reversal (30 bars = 30min)
elif strong_trend:
    signal = -slope(close, 240)           # Fade the trend
else:
    signal = -range_position(close, 120)  # Mean-revert within range
```

---

4. XGBoost Feature Importance (Top 20 -- Table 10)

The most important factors in the XGBoost model (nonlinear factor selection):

Rank	ID	Name	Importance	Category
1	006	VWAP Deviation	6.04%	VWAP
2	061	Alpha101_12_Modern	4.06%	Classic
3	023	Normalized-Momentum TsRank	3.59%	Momentum
4	068	Skewness_Regime_PV_Div	3.55%	**Regime**
5	028	Close-LowxVolume	3.03%	Price range
6	070	Price_Pos_Vol_Interaction	2.27%	Interaction
7	057	TsRank_PV_Divergence	2.26%	Divergence
8	029	High-ClosexVolume	2.15%	Price range
9	018	Range-Position Vol Regime	1.62%	Range
10	045	Kurtosis-Regime Range	1.57%	**Higher moment**

Key insight: Regime-switching factors (#4, #10) and price-range factors (#5, #8) rank highly. For ES single-instrument, the price-range and regime factors are most transferable.

5. Stepwise Selection -- First 5 Factors Provide Most ICIR Gain

| Step | Factor | DeltaICIR |

```
|------|--------|-------|
| 1 | 006 VWAP Deviation | baseline |
| 2 | 046 Vol-Regime Reversal | +0.021 |
| 3 | 079 Regime Vol-Range | +0.050 |
| 4 | 044 Kurtotic Vol Intensity | +0.037 |
| 5 | 041 Price Range Skew | +0.025 |
```

The top 5 stepwise factors combine VWAP, regime-switching, kurtosis, and price-range concepts.

---

6. Complete Factor Library Summary (110 Factors)

Categories Breakdown:

* Regime-switching (IfElse): ~30 factors (031, 034, 036, 038, 042, 045, 046, 049-052, 054-056, 058-060, 064-066, 068-069, 072-073, 076, 079-084, 086-087, 089-090, 094-095, 101-109)
* VWAP-based: ~15 factors (003, 006, 011-014, 016-017, 019, 021, 107-110)
* Price-volume interaction: ~20 factors (004, 007-010, 020, 026-030, 032-033, 035, 037, 039, 043-044, 047-048)
* Amount efficiency: ~12 factors (075, 078, 088, 091-093, 096-099)
* Trend regression (R^2/Slope/Resi): ~8 factors (080-086, 100)
* Higher-order moments (Skew/Kurt): ~8 factors (039-042, 044-045, 063, 074)
* Classic/Other: remaining (~17 factors)

All 110 Formulas Available

Full formulas extracted from Appendix P -- stored in this file for reference. See paper PDF at /tmp/factorminer.pdf for the complete table.

---

7. Next Steps for Implementation

1. Build regime detection module for ES: rolling vol, skew, kurt, R^2
2. Implement Tier 1 factors (046, 042, 080, 045 patterns) adapted for single-instrument
3. Backtest regime-switching vs. static factors on ES 1-min data
4. Test factor combination: start with IC-weighted (simplest), then try XGBoost
5. Validate window sizes: paper uses 10-min bars; need to optimize for ES 1-min or 5-min
6. Key hypothesis: Regime-conditional reversal should work on ES because S&P 500 has well-documented volatility clustering and regime-dependent mean reversion

PFOF & Option Internalization (Ernst & Spatt, RFS 2026)

Paper: Ernst, Thomas and Chester S. Spatt. "Payment for Order Flow and Option Internalization." The Review of Financial Studies, 2026. DOI: 10.1093/rfs/hhaf108
NBER Working Paper: w29883 (March 2022, revised May 2022)
Reviewed: 2026-03-02
Relevance Score: 35 (from Feb 15-26 arxiv q-fin scan)

---

1. Paper Summary

Core Question

How does payment for order flow (PFOF) and internalization work differently in options vs equities, and what are the competitive dynamics?

Key Findings

A. Options PFOF is Much Larger Than Equity PFOF
* In stocks, PFOF is small and retail trades receive meaningful price improvement (especially at minimum spreads)
* In equity options, PFOF is large -- two-thirds of all PFOF comes from options
* This creates a broker conflict of interest: incentive to push customers toward options (higher PFOF asset class)

B. All Options Trade On-Exchange -- But Internalization Still Happens
* Unlike equities (where retail trades go off-exchange to wholesalers like Citadel), ALL option trades execute on-exchange
* But exchange rules facilitate internalization through two key mechanisms:
1. Designated Market Maker (DMM) Assignments -- Exchanges appoint DMMs at the stock-specific level. A DMM can route incoming orders to their own quotes, regardless of position in the price-time or price-pro-rata queue
2. Price Improvement Mechanisms (PIMs) / Auctions -- A market maker brings a retail trade and proposes a price, then other participants can enter the auction and propose better prices

C. The "First Five Contracts" Rule -- A Key Structural Advantage
* If a DMM quotes at the NBBO, they get priority to execute any order of 5 contracts or fewer in full
* This effectively allows the DMM to internalize small retail orders
* This is a mechanism that does NOT exist in equity wholesaling
* Creates a barrier to entry in option wholesaling -- you need DMM status across exchanges to internalize effectively

D. DMMs Who Pay PFOF Give Worse Prices
* Exploiting variation in DMM assignments across exchanges, the paper shows:
- Retail traders receive less price improvement from DMMs who pay PFOF
- They receive worse prices overall from PFOF-paying DMMs
* Imperfectly competitive: DMM allocation rules protect wholesaler profits

E. Market Structure is Concentrated
* The "Big Three" option wholesalers -- Citadel, Susquehanna, and Wolverine -- handle ~85-90% of retail options flow
* Concentration has been rising over time (from ~73% to ~90% by Q2 2021)
* Morgan Stanley and Dash/IMC are other notable PFOF providers

---

2. How PFOF/Internalization Affects Option Pricing and Market Maker Behavior

Pricing Effects

* Wider effective spreads for retail: DMMs paying PFOF provide less price improvement, meaning retail traders systematically pay more
* Cross-subsidy dynamic: The high PFOF in options means market makers earn enough spread to pay brokers for the flow AND still profit -- retail is the product
* Penny tick vs nickel tick: The minimum tick size matters. In nickel-tick options, there's more room for price improvement (and more room for PFOF extraction). Penny-tick options compress this margin

Market Maker Behavior

* Segmentation is key: Market makers can distinguish retail from institutional flow via the PFOF channel. Retail flow is less adversely selected = more profitable to internalize
* DMM status is strategically valuable: Wholesalers actively seek DMM assignments across multiple exchanges specifically to internalize retail flow
* Auction behavior: In PIM auctions, the initiating market maker has information advantage -- they know the order before other participants can bid. Auctions suffer from winner's curse, reducing competition

---

3. Impact on Interpreting Options Flow Data (OPRA/GEX)

Critical Implications for Our GEX Service

A. Not All OPRA Trades Are "Price Discovery" Trades
* A significant chunk of on-exchange option trades are internalized retail orders -- they look like regular exchange prints on OPRA, but they're actually pre-arranged between a wholesaler and a retail broker
* PIM auction trades and DMM-internalized trades will show up in OPRA data as regular trades
* These trades may NOT represent dealer gamma positioning decisions -- they're execution of existing retail orders, not dealer-initiated hedging

B. Signed Trade Classification Issues
* Our GEX model uses signed OPRA trades to infer dealer positioning
* Internalized trades likely appear as trades at or near the NBBO midpoint (price improvement trades), or at the bid/ask
* Standard Lee-Ready or tick-rule signing may misclassify the dealer's role:
- When a wholesaler internalizes a retail buy, the wholesaler is SELLING -> this is dealer short gamma
- But the trade may print at a price that looks like a seller-initiated trade (price improved below ask)
- Net effect: potential bias in our directional signing

C. The "First Five" Rule Creates Systematic Small-Order Bias
* Orders <=5 contracts are systematically internalized by DMMs
* These small orders are disproportionately retail
* On OPRA, these prints look identical to any other trade
* Our GEX calculation may over-weight or misinterpret these small internalized trades

D. Concentration Risk
* 85-90% of retail flow goes through three wholesalers
* When Citadel/Susquehanna/Wolverine internalize, they're accumulating concentrated directional risk that they hedge in the underlying
* Their hedging activity IS what drives GEX effects -- but the timing and method of their hedging may differ from what the raw OPRA flow suggests
* Lag between internalization and delta-hedging could matter for intraday GEX accuracy

---

4. Implications for Our Dealer Positioning Model

What Changes

1. Flow Segmentation Needed: Ideally, we'd separate retail-internalized flow from institutional flow. Some signals:
   - Trade size <=5 contracts (high probability internalized via DMM rule)
   - Trades at improved prices (between NBBO bid/ask and midpoint)
   - Trades flagged through exchange-specific PIM indicators (some exchanges report these differently)

2. Dealer Gamma Sign May Be More Reliable for Large Trades: Large institutional options trades are less likely to be internalized -> more likely to reflect genuine market maker positioning -> better GEX signal

3. Wholesaler Hedging Lag: The dealer who internalizes retail options flow (Citadel etc.) may not hedge instantaneously. They:
   - Aggregate flow
   - Net opposing positions
   - Hedge the residual, often in futures
   - This creates a temporal disconnect between when we see the OPRA print and when the delta hedge actually moves the underlying

4. PFOF-Driven Order Routing May Skew Exchange Distribution: Brokers route to exchanges where their preferred DMM has assignments -> certain exchanges see concentrated retail flow -> exchange-level data becomes less representative of "true" market sentiment

Actionable Improvements

* Filter by trade size: Weight trades >10 contracts more heavily in GEX calculations (less likely internalized)
* Track trade-at-midpoint ratio: High ratio -> more internalization activity -> less reliable directional signal
* Consider adding a lag factor: Account for the 1-5 minute delay between internalization print and actual delta hedge
* Monitor exchange distribution: Sudden shifts in exchange-level volume may indicate routing changes, not sentiment changes

---

5. Bottom Line

The paper confirms that a significant portion of options trades visible on OPRA are pre-arranged retail internalization, not organic price discovery. This matters for GEX because:

* Our signed trade approach may have systematic biases from internalized flow
* Small trades (<=5 contracts) are almost certainly retail internalization -> low information content for dealer positioning
* The actual dealer hedging happens with a lag and in aggregate, not trade-by-trade
* The three dominant wholesalers' net positioning drives the real GEX effect, but we only see their gross trades

Recommendation: Implement a trade-size filter (weight down <=5 contract trades) and consider a price-improvement filter to improve GEX signal quality. The paper suggests the most informative trades for dealer positioning are larger institutional trades that go through normal exchange matching.

Related Papers

* Bryzgalova, Pavlova, Sikorskaya (2023). "Retail Trading in Options and the Rise of the Big Three Wholesalers." *Journal of Finance*. -- Complementary paper on wholesaler market share, retail identification flags, and concentration dynamics.
* Ernst, Spatt, Sun (2023/2025). "Would Order-By-Order Auctions Be Competitive?" *Journal of Finance*. -- Models proposed regulatory changes to the auction system.

Metaorder Flow & Dealer Hedging

Paper: "Metaorder modelling and identification from public data" (arXiv:2602.19590v1)
Authors: Ezra Goliath, Tim Gebbie (University of Cape Town)
Date: Feb 23, 2026
Relevance Score: 30 (from Feb 15-26 arxiv q-fin scan)
Date Reviewed: 2026-03-02

Core Question: What Drives Long-Range Correlation in Order Flow?

The LMF (Lillo-Mike-Farmer) Theory

The paper validates a fundamental result in market microstructure: trade sign autocorrelations exhibit long-range power-law decay, and this is caused by order-splitting behavior (metaorders).

Key relationship (the LMF law):
```
? = ? - 1
```

Where:
* ? = decay exponent of the autocorrelation function C(?) ? ?^(-?) of trade signs (? = ±1)
* ? = power-law exponent of metaorder length distribution P(L) ? L^(-?-1)

Translation: The reason order flow is predictable (long-memory) is that institutional traders split large parent orders into many small child orders executed over time. The distribution of these metaorder lengths (heavy-tailed, power-law) directly determines how slowly the autocorrelation decays.

How Metaorders Create Predictable Patterns

1. Order splitting: A fund wanting to buy 100k shares splits into hundreds/thousands of small child orders
2. Persistence: Each child order has the same sign (+1 for buy), creating runs of same-sign trades
3. Mixing: Many such metaorders from different traders overlap in time
4. Aggregation: The aggregate trade-sign series inherits long memory from the power-law distribution of metaorder lengths
5. Predictability paradox: Order flow is predictable (long-range correlated) even though price changes are not -- this is because price impact is concave (square-root law), so the predictable order flow doesn't translate into predictable price changes

Key Stylized Facts Validated

1. Square Root Law (SQL): Price impact I(Q) = Y x sigma_D x sqrt(Q/V_D) -- impact is concave in order size
2. Time independence: Metaorder impact doesn't depend on execution duration T
3. Concave execution profile: I(?xQ) ? sqrt? x I(Q) during execution
4. Post-execution decay: Impact partially reverts after metaorder completion, with ? ~= 0.22-0.24

Methodological Innovation

The paper's key contribution: you don't need proprietary broker codes to reconstruct metaorders. Using publicly available TAQ data + a synthetic trader assignment algorithm, they reproduce all stylized facts. This was validated on JSE (Johannesburg) data for 100 stocks over 3 years.

The algorithm:
* Assume N effective traders (they tested N=50, N=150)
* Assign trades to synthetic traders using power-law participation weights

* Define metaorders as consecutive same-sign trades from a single synthetic trader
* Results are reasonably robust to choice of N and distribution shape

---

Assessment: Can We Detect Metaorders in ES from 1-Min Data?

What We Have

* ES 1-min bars (OHLCV + volume)
* Signed OPRA trade data (options flow)

Feasibility: **Partially -- but with major limitations**

What WORKS with our data:

1. Trade-sign autocorrelation measurement: We can compute C(?) from tick-level ES trade data (if we have it) or approximate from 1-min signed volume. The ACF decay exponent ? tells us about metaorder presence.
2. Volume clustering analysis: 1-min volume patterns can reveal institutional execution footprints -- elevated volume persisting across multiple bars in a directional pattern.
3. OPRA signed flow: Options flow signatures could reveal hedging-driven metaorders. Large delta-hedging flows from dealer positioning create metaorder-like patterns in ES.

What DOESN'T work:

1. Trade-by-trade reconstruction is impossible from 1-min bars -- we lose individual trade assignment needed for the Maitrier et al. algorithm
2. No broker codes -- even with tick data, we can't identify individual traders on CME
3. ES is a single instrument -- the paper used 100 stocks, which helps average noise. We have one contract.

Practical Application Angles for ES Trading

1. Order Flow Autocorrelation as a Regime Indicator (FEASIBLE)

* Compute rolling autocorrelation of signed 1-min volume (or trade imbalance)
* High autocorrelation -> metaorder-driven market -> trend persistence likely
* Low autocorrelation -> noise/information-driven -> mean reversion more likely
* This is essentially a regime filter -- when ? is low (strong long-memory), momentum signals should work better

2. Volume Profile Analysis (FEASIBLE)

* Track cumulative directional volume over rolling windows (e.g., 30-60 min)
* Persistent one-sided flow above threshold -> likely metaorder execution in progress
* The square-root impact law tells us: impact ? sqrt(Q/V_D), so we can estimate expected impact from observed flow
* When observed price move exceeds sqrtQ prediction -> overreaction, likely to revert
* When observed price move is below -> metaorder may still be in progress

3. Post-Execution Decay Trading (SPECULATIVE)

* If we detect a likely metaorder completion (sudden drop in directional volume after sustained flow), the paper shows impact decays with ? ~= 0.22
* This implies partial mean reversion after large one-sided flows stop
* Timeframe: the decay happens over z = t/T where T is the metaorder duration

4. OPRA Cross-Signal (MOST PROMISING)

* Large option trades trigger dealer delta-hedging -> creates metaorder-like flows in ES
* Our signed OPRA data could predict upcoming ES metaorder flow before it fully materializes
* Sequence: big option trade -> dealer needs to hedge -> systematic buying/selling in ES futures over next N minutes
* This is a leading indicator of metaorder activity

Concrete Signal Ideas

Signal 1: Autocorrelation Regime Filter

```
acf_30 = rolling_autocorrelation(signed_volume, lag=1, window=30_bars)
if acf_30 > threshold:
    # Strong metaorder regime -> favor momentum signals
else:
    # Noise regime -> favor mean reversion
```

Signal 2: Flow Persistence Indicator

```
cum_signed_vol = rolling_sum(signed_volume, window=30)
normalized_flow = cum_signed_vol / sqrt(rolling_avg_daily_volume)
```

Signal 3: Impact Residual

```
expected_impact = Y  sigma_D  sqrt(abs(cum_flow) / V_D)
actual_impact = price_change_over_window
residual = actual_impact - expected_impact
```

Signal 4: OPRA-to-ES Flow Prediction

```
large_option_trades = filter(opra_trades, abs(delta_notional) > threshold)
predicted_es_hedge_flow = sum(dealer_delta_hedge_needed)
```

Bottom Line Assessment

Aspect	Rating	Notes
Theoretical insight	?????	Fundamental result connecting micro to macro
Direct applicability to ES	?????	We lack tick-by-tick trade data for reconstruction
Derived signal potential	?????	Regime filter + flow persistence very promising
OPRA cross-signal	?????	Best angle -- option flow predicts hedging metaorders
Backtestability	?????	Flow persistence and ACF signals testable from 1-min bars

Recommendation

* Don't try to reconstruct metaorders from 1-min ES data -- not enough granularity
* DO build: (1) autocorrelation regime filter, (2) flow persistence indicator, (3) impact residual signal
* Best opportunity: Use OPRA signed flow to predict upcoming hedging-driven metaorders in ES
* The square-root impact law gives us a calibrated expectation for how much price should move given observed flow -- deviations from this are tradeable

Key Papers to Follow Up

* Maitrier, Loeper, Bouchaud (2025) -- "Generating realistic metaorders from public data" (arXiv:2503.18199) -- the core methodology paper
* Sato & Kanazawa (2023) -- Original LMF validation on Tokyo Stock Exchange (PRL 131, 197401)
* Bouchaud (2025) -- "The universal law behind market price swings" (Physics 18, 196) -- accessible overview
* Tóth et al. (2011) -- "Anomalous price impact and the critical nature of liquidity" -- the square-root law paper

Multi-Agent LLM Trading Teams (arXiv 2602.23330)

Paper: "Toward Expert Investment Teams: A Multi-Agent LLM System with Fine-Grained Trading Tasks"
Authors: Kunihiro Miyazaki, Takanobu Kawahara, Stephen Roberts, Stefan Zohren (Oxford)
arxiv: 2602.23330v1 (Feb 26, 2026)
Relevance Score: 51

TL;DR

Fine-grained task decomposition for LLM trading agents (telling them exactly what to analyze, not just "analyze this") significantly improves Sharpe ratios vs coarse-grained role assignments. The Technical Analysis agent is the primary performance driver. Most other agents (fundamental, qualitative, macro) actually introduce noise. The system's outputs are lowly correlated with the market index (~0.4), creating diversification value.

Architecture

3-Level Hierarchy (Bottom-Up)

```
Level 1: Analyst Agents (4 specialists per stock)
  +-- Technical Agent -> Score [0-100] + rationale
  +-- Quantitative Agent -> Score [0-100] + rationale
  +-- Qualitative Agent -> Score [1-5] + rationale
  +-- News Agent -> Score [1-5] + rationale

Level 2: Aggregation (2 agents)
  +-- Sector Agent -> Re-scores vs sector averages -> Score [0-100]
  +-- Macro Agent -> Regime assessment (5 dimensions) -> Scores [0-100]

Level 3: Portfolio Manager Agent
  +-- Synthesizes bottom-up (Sector) + top-down (Macro) -> Final score [0-100]
```

Key Design Choices

* Market-neutral long/short portfolio (equal weight long + short)
* Monthly rebalancing on TOPIX 100 (Japan large-cap)
* GPT-4o with knowledge cutoff Aug 2023; backtest Sep 2023 - Nov 2025 (27 months)
* Temperature = 1.0 (preserves ensemble diversity; median of 50 trials used)
* Strict leakage prevention: only data available at decision time

Fine-Grained vs Coarse-Grained (The Core Finding)

Fine-grained = pre-calculate indicators, give agents structured inputs:

* Technical: RoC (8 timeframes), Bollinger Z-score, MACD, RSI, KDJ
* Quant: ROE, ROA, P/E, EV/EBITDA, D/E, etc. organized into 5 dimensions

Coarse-grained = dump raw data and let the LLM figure it out:

* Technical: raw daily prices for 1 year
* Quant: raw financial statement line items

Result: Fine-grained wins significantly ($p<0.0001$ for portfolio sizes 20-50 stocks)

Performance Results

Sharpe Ratios (median, fine-grained, all agents)

Portfolio Size	Fine-Grained SR	Coarse-Grained SR
N=10	0.54	0.66
N=20	0.84	0.65
N=30	0.84	0.76
N=40	0.79	0.62
N=50	0.90	0.63

Ablation Findings (What Actually Matters)

Technical Agent is king. Removing it tanks performance in fine-grained setting:

* w/o Technical at N=50: SR drops by -0.66 (massive)

Most other agents ADD NOISE:

* Removing Quantitative, Qualitative, News, or Macro agents *improves* performance in fine-grained setting
* This suggests "more agents != better" -- signal-to-noise matters

Exception -- News in coarse-grained: News agent is beneficial in coarse mode (removing it hurts by -0.57 to -0.75), but not in fine-grained. Interpretation: when technical signals are weak (coarse), news compensates.

Information Propagation

* Fine-grained prompts increase semantic similarity between Technical Agent and Sector Agent outputs (cosine sim +0.022)
* This means technical insights actually reach the PM's decision
* Without fine-grained prompts, the system defaults to fundamental-analysis-biased reasoning

Portfolio Optimization

* Agent composite has only ~0.4 correlation with TOPIX 100 index
* Blending 50/50 with index beats both standalone components on Sharpe
* Net of 10bps transaction costs, still attractive

What We Can Steal for Our System

1. **Pre-compute indicators, don't let LLMs do math** ???

The biggest takeaway. When we feed signals to our aggregator, we should pre-calculate everything: normalized indicators, z-scores, ratios. Don't dump raw OHLCV and expect the LLM to compute RSI correctly. This alone drove the performance gap.

Action: Ensure all signals fed to our system are pre-processed, normalized, and structured. No raw data dumps.

2. **Technical signals dominate** ??

For their Japanese stock universe, technical analysis was the primary alpha driver. Fundamental/qualitative/news agents mostly added noise. This challenges the "more diverse signals = better" assumption.

Caution: This was monthly rebalancing on large-cap Japanese stocks. For ES futures intraday, the signal hierarchy may differ. But it's a warning: don't assume every signal source helps. Measure ablation impact.

Action: Run ablation tests on our signal sources. Which ones actually improve vs degrade?

3. **Hierarchical aggregation with sector/regime adjustment** ??

Their Sector Agent re-scores stocks against sector peers. Their Macro Agent provides regime context. The PM merges bottom-up + top-down views.

Relevance to us: We could add a regime-aware layer to our signal aggregator -- e.g., different weighting of momentum vs mean-reversion signals based on a macro/volatility regime assessment.

4. **Ensemble via diversity (multiple agent configs)** ??

They create 6 strategy variants (all-agents + 5 leave-one-out) and combine via equal risk contribution. The heterogeneity itself is valuable because outputs are somewhat uncorrelated.

Action: If we run multiple signal aggregation configs, blending their outputs could reduce variance.

5. **Information propagation matters** ?

Not just what agents say, but whether downstream agents actually absorb the signal. They measured this with embedding cosine similarity. Fine-grained prompts improved propagation of technical signals.

Action: When designing multi-step LLM pipelines, verify that intermediate outputs actually influence final decisions. Use structured output formats that force downstream agents to reference upstream findings.

6. **Natural language intermediaries enable interpretability** ?

All agent outputs are text + scores. This makes the system auditable -- you can read why the system chose a position. Valuable for debugging bad trades.

Limitations & Caveats

1. Monthly rebalancing on Japanese stocks -- very different from intraday ES futures
2. 27-month backtest -- too short to confirm robustness across regimes
3. GPT-4o only -- no comparison with other models
4. No transaction cost sensitivity beyond 10bps one-way
5. The "fine-grained" advantage might be feature engineering, not task decomposition -- pre-computing RSI vs feeding raw prices is basically just better data prep. The paper somewhat conflates prompt design with feature engineering.
6. Ablation paradox: if removing 4 of 5 specialist agents improves performance, maybe the system should just be a Technical Agent + PM. The paper doesn't explore this simpler architecture.

Bottom Line for Us

Practical rating: 6/10 -- The core insight (pre-compute features for LLMs, don't make them do math) is genuinely useful and we should verify we're doing this. The multi-agent architecture itself is over-engineered for what it delivers -- most agents hurt performance. The ablation results are actually the most valuable finding: they tell you to be ruthless about signal selection.

For our ES futures trading: pre-compute all technical indicators, structure them clearly for LLM consumption, and measure whether each signal source actually helps. Don't add agents/signals for the sake of diversity.

Reviewed: 2026-03-02

VOLARE Volatility Database

Paper: arXiv:2602.19732v1 (Feb 23, 2026)
Authors: Cipollini, Cruciani, Gallo, Insana, Otranto, Spagnolo (Florence/Messina/NYU/Sapienza)
Platform: https://volare.unime.it
Score: 30 (from Feb 15-26 q-fin scan)

What VOLARE Provides

An open-access database of standardized daily realized volatility and covariance measures computed from ultra-high-frequency (tick-level) financial data. Replaces the discontinued Oxford-Man Realized Library (died mid-2022).

Asset Coverage

Class	Assets	Data Start	Source Format
Stocks	40 US equities (all Dow 30 + 10 from S&P 100)	2015-01-02	tickms (millisecond)
FX	5 pairs (AUDUSD, EURUSD, GBPUSD, USDCAD, USDJPY)	2009-09-25	tickbidask
Futures	5 contracts: **ES (E-mini S&P 500)**, CL (Crude), NG (NatGas), GC (Gold), C (Corn)	2009-09-28	tickbidask

Key: ES futures are included -- data from Sep 2009 onward.

Realized Measures Computed (Daily)

Univariate (10 measures, at 1-min and 5-min sampling):
1. Realized Variance (rv) -- rv1, rv5, rv5_ss (subsampled)
2. Bipower Variation (bv) -- bv1, bv5, bv5_ss -- separates continuous from jump component
3. Realized Semivariances (rsp, rsn) -- positive/negative decomposition of variance
4. Realized Quarticity (rq) -- measures tail risk / variance of variance
5. Realized Kernel (rk) -- noise-robust estimator using 1-second data
6. Parkinson Range (pr) -- daily high-low based
7. Garman-Klass Range (gkr) -- OHLC-based
8. Realized Range (rr) -- intraday range-based
9. Median Realized Variance (medrv) -- jump-robust
10. Minimum Realized Variance (minrv) -- jump-robust

Multivariate (3 measures):
* Realized Covariance (rcov)
* Realized Bipower Covariance (rbpcov)
* Realized Semicovariances (rscov) -- concordant/discordant decomposition

How Realized Vol Is Calculated

1. Raw data source: Kibot tick data (millisecond-precision for stocks, second for futures)
2. Cleaning: Brownlees-Gallo (2006) outlier detection for stocks; mid-quote prices for FX/futures (no separate cleaning needed)
3. Sampling: Tick data aggregated to regular intervals (1-min, 5-min) using last-available-price
4. RV formula: Standard sum of squared intraday returns: RV = ?(r_k^2)
5. Subsampling: For 5-min measures, computes on 5 shifted 1-min grids and averages -> reduces variance/bias
6. Minimum threshold: Only computed on days with >=40 observations or >=2 hours of trading
7. Trading hours:
   - Stocks: 9:30-16:05 ET (extra 5 min to capture closing trades)
   - Futures: Sun 18:00 - Fri 17:00 ET (with daily 17:00-18:00 maintenance break)
   - FX: Sun 17:00 - Fri 17:00 ET

Built-in Models

VOLARE also provides real-time model estimation on the platform:

* HAR (Heterogeneous Autoregressive) -- daily/weekly/monthly RV components
* HAR-Q -- HAR augmented with realized quarticity
* MEM (Multiplicative Error Model) -- Engle-Gallo
* AMEM (Asymmetric MEM) -- captures leverage effects
* Forecasting with rolling/expanding windows

Assessment: Relevance to Our Vol Regime Detector

Current Setup

We use VIX level + VIX term structure ratio (VIX/VIX3M) for vol regime classification:

* Low vol: VIX < 15, contango
* Normal: VIX 15-25
* High vol: VIX > 25, backwardation
* Crisis: VIX > 35

What VOLARE Adds

? STRONG POTENTIAL -- ES futures realized vol is directly available (since 2009)

Specific Value-Adds:

1. Realized vs Implied Divergence Signal
   - VIX = implied vol (30-day forward-looking, option-derived)
   - VOLARE RV = realized vol (backward-looking, from actual price movements)
   - Spread between VIX and RV is a known predictor -- the "variance risk premium"
   - When VIX >> RV: market pricing in fear that hasn't materialized -> potential mean reversion
   - When VIX ~= RV or VIX < RV: realized vol catching up -> regime shift confirmation

2. Semivariance Decomposition (Unique)
   - VOLARE provides RSP (positive semivariance) and RSN (negative semivariance) separately
   - RSN/RSP ratio = directional vol asymmetry
   - Rising RSN relative to RSP = downside vol dominating -> earlier regime shift detection than symmetric VIX
   - This is not available from VIX at all

3. Jump Detection via Bipower Variation
   - BV captures continuous component; RV - BV = jump component
   - Spikes in jump component signal regime transitions
   - Our VIX-based detector is slow to catch jump-driven regimes

4. Realized Quarticity = Tail Risk Gauge
   - RQ measures "variance of variance" -- how unstable is vol itself
   - High RQ = volatile volatility -> regime instability
   - Could serve as a confidence/stability measure for our regime labels

Practical Integration Path:

```
Enhanced Regime Detector = f(VIX_level, VIX_term_structure, RV_ES, VRP, semivar_ratio, jump_component)
```

Where:

* `VRP = VIX^2 - RV_ES` (variance risk premium)

* `semivar_ratio = RSN / (RSN + RSP)` (downside fraction)
* `jump_component = RV - BV` (jump intensity)

?? Limitations

1. Daily frequency only -- our intraday trading needs real-time signals; VOLARE provides end-of-day realized measures
2. Lagging indicator -- RV is backward-looking by construction; VIX is forward-looking
3. Not a replacement but a complement -- best used to confirm/validate VIX-based regime calls, not replace them
4. Data access uncertainty -- need to verify API/bulk download actually works and is timely (site is new)
5. Limited futures coverage -- only ES, CL, NG, GC, C (no NQ, no bonds)
6. Academic project -- may not update reliably; the Oxford-Man library died too

? Bottom Line

VERDICT: Worth integrating as a secondary signal, not a primary replacement.

The variance risk premium (VIX^2 - ES realized variance) is the highest-value addition. It's a well-documented predictor in the academic literature, and VOLARE provides clean ES realized variance going back to 2009. The semivariance decomposition is also genuinely novel for our regime detector.

Recommended next step: Download ES futures RV data from VOLARE and backtest whether adding VRP and semivariance ratio improves regime transition detection timing vs our current VIX-only approach. Key question: does VRP compress before regime transitions in a way that's actionable for intraday trading?

Priority: MEDIUM -- Useful enhancement but not urgent. Our VIX-based detector works for coarse regime buckets. This would add nuance, especially around regime transitions.

Cross-Asset Spillovers via SDF

Paper: Avramov & He (2026) -- arxiv 2602.20856v1
Date reviewed: 2026-03-02
Relevance score: 54 (from research scan)

TL;DR for Daniel

Bottom line: Interesting academic framework, but NOT directly usable for intraday ES trading. The paper operates on monthly stock return data using 138 firm-level fundamental signals. It's a cross-sectional equity model, not a cross-asset (stocks->futures) lead-lag model. The core insight -- that large, low-turnover firms are "net transmitters" of predictive information to smaller stocks -- is intellectually useful but the timescale and asset universe don't map to your intraday ES setup.

However, the concept of building a cross-asset spillover matrix (?) is powerful and could be adapted to an intraday multi-asset framework using our existing data.

Paper Methodology

What They Do

1. Framework: Joint estimation of signal weights (? vector) and a cross-asset connection matrix (?) by maximizing the portfolio Sharpe ratio
2. Signal matrix S: N assets x M signals (138 firm-level characteristics like size, value, profitability, investment, momentum)
3. Trading strategy: ? = ?' S' ? -- portfolio weights are a function of all assets' signals, not just own signals
4. Connection matrix ?: NxN matrix where ?(i,j) = predictive influence of asset i's signals on asset j's returns
5. Estimation: Iterative ridge regression (Britten-Jones approach), 5-fold CV for ? selection, rolling 120-month windows

Key Innovation

Traditional models: asset i's own characteristics -> asset i's returns (self-prediction, diagonal ?)
This paper: asset i's characteristics -> asset j's returns (cross-prediction, full ? matrix)

The off-diagonal elements of ? carry MORE information than diagonal (self-predictive) elements: avg |off-diag| = 0.0805 vs avg |diag| = 0.0068 in the toy example.

Performance (Out-of-Sample, 1973-2023)

* Cross-predictive SR-max: Sharpe 2.21 (spread portfolios), 3.32 (bi-sort portfolios)
* Self-predictive SR-max: Sharpe 1.42 (spread), 2.06 (bi-sort) -- notably worse
* Expected-return max: Sharpe 0.45 -- terrible risk-adjusted performance
* Alpha vs all major factor models: ~0.25%/month, t-stat > 11

Net Transmitters vs Net Receivers (Diebold-Yilmaz Network Analysis)

Using the ? matrix, they aggregate rows/columns to compute directional spillover:

Net Transmitters (information leaders):
* Large-cap, low-turnover stocks
* Typically mature, fundamentals-rich firms
* Their signals predict returns of OTHER stocks

Net Receivers (information followers):
* Small-cap, high-turnover stocks

* Value-oriented, high profitability, low investment, strong past returns
* Their returns are predicted BY other stocks' signals

Signal Categories That Matter Most

High weight in ? (matter a lot):

* Investment signals (asset growth, capex)
* Value signals (dividend yield, book-to-market)
* Profitability signals (ROE, operating profit)

Low weight in ? (don't matter much):

* Momentum
* Short-term reversal
* Seasonality

This is notable: fundamentals-based cross-prediction dominates return-based cross-prediction.

Post-2000 Decay

Performance declined notably after 2000 (echoing the broader anomaly decay literature). 5-year trailing Sharpe ~1.2 by end of 2023, still beating all individual factors but far from the 1990s glory days.

Can We Use This for Intraday ES Trading?

Direct Application: NO ?

Mismatches:
1. Timescale: Monthly rebalancing. We need 1-min to 1-hour signals.
2. Asset universe: 138+ sorted equity portfolios. We trade ES futures.
3. Signal type: Firm fundamentals (ROE, asset growth). We need intraday price/flow signals.
4. Estimation window: 10-year rolling. Useless for intraday regime changes.
5. Cross-section focus: The paper exploits cross-sectional variation among stocks. We need time-series prediction of a single instrument (ES).

Adapted Application: MAYBE ??

The concept of a cross-asset spillover matrix is powerful. We could build an intraday version:

Our "? matrix" -- Cross-Asset Lead-Lag at Hourly Frequency:

Assets in our universe (hourly data available):

* ES (target to predict)
* NQ (tech/growth proxy)
* NVDA (AI/semiconductor bellwether)
* TLT (duration/risk-off)
* DXY (dollar strength)
* GC (gold/safe haven)
* VIX (implied vol)
* HYG (credit risk)
* USDJPY (carry trade)

Proposed signals (per asset, per bar):

* Returns (1h, 2h, 4h lookback)
* Relative volume (vs 20-period avg)

* Volatility ratio (realized vs recent avg)
* Momentum (deviation from intraday VWAP-like level)

The question: Do NQ returns, NVDA volume spikes, TLT moves, or VIX changes at hour H predict ES at hour H+1?

This is essentially a VAR(p) or Granger-causality framework -- simpler than the paper's SDF approach but captures the same lead-lag concept.

Practical Backtest Design (If We Proceed)

Simplified Cross-Asset Lead-Lag Test

Hypothesis: Hourly returns of NQ, NVDA, TLT, VIX, DXY, GC lead ES returns by 1 hour, and the lead-lag structure is time-varying.

Data available:

Asset	File	Freq	Rows
ES	es_1h_tv.csv	1h	5001
NQ	nq_1h_tv.csv	1h	~5000
NVDA	nvda_1h_tv.csv	1h	~5000
TLT	tlt_1h_tv.csv	1h	5001
VIX	vix_1h_tv.csv	1h	~5000
DXY	dxy_1h_tv.csv	1h	~5000
GC	gc_1h_tv.csv	1h	~5000
SPX	spx_1h_tv.csv	1h	~5000

Method:
1. Compute hourly returns for all assets
2. Rolling 60-bar window: estimate simple ? as OLS coefficients of ES_ret(t+1) on [ES_ret(t), NQ_ret(t), NVDA_ret(t), TLT_ret(t), VIX_ret(t), DXY_ret(t), GC_ret(t)]
3. Track which coefficients are persistently significant -> identifies "transmitters" to ES
4. Generate signal: weighted combination of lagged returns using rolling ? estimates
5. Backtest: long ES if signal > threshold, short if signal < -threshold

Checklist items (per BACKTEST_CHECKLIST.md):
* [ ] IS/OOS split: first 60% IS, last 40% OOS
* [ ] No lookahead: only use data from previous bars
* [ ] Intraday actionable: signal available at bar close, trade next bar
* [ ] t-stat >= 2.0 required
* [ ] Minimum 30 trades per direction

Expected Challenges

1. Correlation != lead-lag: ES and NQ are ~95% correlated contemporaneously, making it hard to isolate leading information
2. Non-stationarity: Lead-lag relationships shift across regimes
3. Signal decay: Any hourly-frequency alpha gets arbitraged fast
4. Our data starts 2023-2025: Only ~2 years of hourly data, may not be enough for robust estimation

---

Verdict

Criterion	Assessment
Academic quality	High -- rigorous framework, solid OOS results
Novelty	Moderate -- extends existing SDF/factor literature with cross-prediction
Direct applicability to ES intraday	**LOW** -- wrong timescale, wrong asset class

Conceptual value	**MODERATE** -- the ? spillover matrix idea is portable
Worth backtesting adapted version?	**MAYBE** -- low priority, would need careful data alignment
Expected edge if it works	Small -- likely 5-10 bps/trade at best, degrades fast

Recommendation

Don't prioritize this. The paper is about cross-sectional equity predictability at monthly frequency -- a completely different game from intraday ES futures. The one useful takeaway is the idea of systematically measuring and exploiting lead-lag relationships across assets using a connection matrix. But we can implement that idea with simpler tools (rolling VAR, Granger causality) that are better suited to our intraday setup.

If Daniel wants to pursue the cross-asset lead-lag concept, a more productive path would be:
1. Start with simple Granger causality tests between NQ->ES, NVDA->ES, TLT->ES at hourly frequency
2. If any show significance, build a rolling prediction model
3. The paper's key insight to retain: off-diagonal (cross-asset) predictability often exceeds on-diagonal (self-predictability). If NQ's move predicts ES better than ES's own lagged returns, that's a signal.

Filed: 2026-03-02 | Paper: Avramov & He (2026) | Priority: Low for direct use, moderate for conceptual inspiration

RORO Index (Chari et al., NBER WP 31907)

Source

* NBER WP 31907
* Data: https://www.kansascityfed.org/data-and-trends/risk-on-risk-off-index/
* Downloaded daily data to `data/roro_daily.xlsx` (5,669 rows, 2003-05-09 to 2025-03-12)

Construction

First principal component of daily changes in 14 standardized financial variables.
Positive RORO = Risk-Off. Negative RORO = Risk-On.

Variables & PCA Loadings (eigenvector weights)

Credit Spreads (highest loadings -- dominate the index)

Variable	Loading	Source
US High Yield OAS (ICE/BoA)	0.388	FRED: BAMLH0A0HYM2
Euro High Yield OAS	0.314	FRED: BAMLHE00EHYIOAS
US BAA - 10Y	0.223	FRED: BAA10Y

Equity & Implied Vol

Variable	Loading	Source
STOXX 600 return	0.416	Bloomberg
S&P 500 return	0.369	Bloomberg: SPX
VSTOXX	0.386	Bloomberg: V2X
VIX	0.352	CBOE
AE MSCI return	0.219	Bloomberg: MSDLEAFE

Funding Liquidity (weak loadings)

Variable	Loading	Source
LIBOR - 3mo OIS	0.146	ICE
TED spread	0.120	ICE/FRB
G-spread (2Y,5Y,10Y)	0.020	WSJ
Bid-ask spread	-0.015	Bloomberg

Gold & Currency (weak loadings)

Variable	Loading	Source
Dollar vs AE currencies	0.183	Fed: DTWEXBGS
Gold price	-0.044	Bloomberg: XAU

Key Insights for ES Trading

1. Credit spreads and equity vol dominate -- HY OAS + VIX/VSTOXX are ~70% of the signal
2. Liquidity barely matters except during systemic crises (2008, COVID)
3. Gold and currency are noise in the PCA -- loading < 0.2
4. Right-skewed, fat-tailed -- risk-off events are bigger than risk-on
5. Outperforms VIX alone for predicting fund flows and returns

Columns in Daily Data

* t: date
* z_spreads: credit spread sub-index
* z_equities: equity/vol sub-index
* z_liquidity: funding liquidity sub-index
* z_goldcurrency: gold/currency sub-index
* z_roro: headline composite RORO index

Implications for Our Work

* The Pine Script versions (v1.5, v3.0, v3.1) use equal-weight z-scores -- academically inferior to PCA
* Our backtests already confirmed HYG/credit and gold/copper are weak for ES prediction
* The academic RORO's strength comes from HY credit + equity vol -- which maps to VIX (already our best signal)
* We now have 22 years of daily RORO data to backtest against ES!

RORO Composite Implementation

Source

* ChatGPT/ScholarGPT conversation (Feb 2026)
* Primary paper: Chari, Stedman & Lundblad (2023) "Risk-On Risk-Off" NBER Working Paper 31907
* Official daily RoRo data available at: anushachari.weebly.com/roro.html

Academic Methodology (Chari et al.)

* PCA-based (first principal component), NOT simple averaging
* 4 risk categories: Credit, Equity, Funding Liquidity, Currency/Gold
* Z-score normalized daily changes, positive = risk-off
* Formula: RoRo_t = First Principal Component of {Z_t(1), ..., Z_t(n)}
* Predictive model: DeltaS&P_{t+h} = ? + ?·RoRo_t + ?_{t+h} where ? < 0

Components Used in Academic Version

1. Credit Risk: ICE BofA BBB OAS (US+Euro), BAA-10Y spread
2. Equity Risk: Inverse SPX/STOXX/MSCI returns, VIX, VSTOXX
3. Funding Liquidity: G-spreads (2y/5y/10y), TED spread, LIBOR-OIS, Treasury bid-ask
4. Currency/Gold: Trade-weighted USD, Gold

TradingView Pine Script v1.5 (built in ChatGPT conversation)

* 13 components, equal-weighted average (not PCA)
* US: SPX, VIX, Gold, DXY, HYG/LQD, Copper, TEDRATE, BAMLH0A0HYM2, DGS2/5/10
* Euro: VSTOXX, IT10Y-DE10Y spread
* Z-scored over 60 days, smoothed with 5-day SMA
* ~90% structural parity with academic version
* Missing: PCA weighting, LIBOR-OIS, Euro BBB OAS, Treasury bid-ask

Our Backtesting Results (what actually works vs doesn't)

Components that FAILED in our OOS testing:

* HYG/LQD credit spread -- no lead-lag on daily bars
* Copper -- no divergence signal
* Gold -- flipped between train/OOS periods
* Treasury yields -- TLT signal unstable

Components that WORK:

* VIX9D/VIX ratio momentum (best intraday signal, 64% WR hourly)
* JPY (yen) -- 71% WR OOS but small sample (17 trades)
* NVDA divergence -- 60-75% WR confirmed OOS
* BTC weekend -- 67-76% WR confirmed OOS
* VVIX extremes -- contrarian buy at >130

Key Insight

The academic RORO uses PCA which auto-weights assets by their covariance contribution.
Equal-weight averaging (as in Pine Script) includes noise from non-predictive assets.
A PCA version using only our validated signals should outperform both the academic version

(which includes noisy assets) and the equal-weight version.

TODO

* [ ] Pull actual Chari RoRo index from anushachari.weebly.com/roro.html and backtest vs ES
* [ ] Build PCA-based composite using only validated signals
* [ ] Add to trading dashboard as "Risk Regime" gauge

Appendix: arXiv q-fin Scan Results

2026-03-02

1. TradeFM: A Generative Foundation Model for Trade-flow and Market Microstructure

* Link: https://arxiv.org/abs/2602.23784
* 524M-parameter Transformer trained on billions of trade events across 9K+ equities. Uses scale-invariant features and universal tokenization to learn order flow patterns. Reproduces heavy tails, vol clustering, and absence of return autocorrelation. 2-3x lower distributional error than Hawkes baselines. Generalizes zero-shot to out-of-distribution markets.
* TRADEABLE IDEA: The scale-invariant features they developed (normalizing trade size, inter-arrival times, price impact) could be replicated as real-time ES order flow features. Their tokenization scheme for encoding trade sequences could feed into a simpler LSTM/transformer for predicting short-term ES price moves from tick data.
* Difficulty: Hard

2. Overreaction as Momentum Indicator in Algorithmic Trading

* Link: https://arxiv.org/abs/2602.18912
* Uses ML (XGBoost, BiLSTM) to predict intraday overreactions from volatility-normalized returns + Twitter emotion features on AAPL. Key finding: ML models beat behavioral benchmarks at ultra-short horizons (1-5 min), while classical momentum dominates at ~10 min. SHAP analysis shows volatility and fear/sadness drive predicted overreactions.
* TRADEABLE IDEA: Build an overreaction detector for ES: flag when 1-5 min returns exceed 2-3 sigma of recent vol, combined with sentiment spikes (X/Twitter fear keywords). Fade extreme moves at 1-5 min, ride momentum at 10 min. The vol-normalization approach is key -- raw return thresholds don't work.
* Difficulty: Medium

3. Stochastic Discount Factors with Cross-Asset Spillovers

* Link: https://arxiv.org/abs/2602.20856
* Jointly estimates firm-level predictive signals AND cross-asset spillovers by maximizing Sharpe ratio. Finds large, low-turnover firms are net information transmitters. Outperforms self-predictive benchmarks across market states. Uncovers directional predictive influence across assets.
* TRADEABLE IDEA: For ES, monitor order flow and price action in mega-cap bellwethers (AAPL, MSFT, NVDA) as leading indicators. The paper confirms large firms transmit information to the broader market. Build a real-time "bellwether momentum" signal: when top-5 holdings by weight move aggressively in one direction, ES follows with a lag.
* Difficulty: Easy

4. Schrödinger Bridges with Jumps for Time Series Generation

* Link: https://arxiv.org/abs/2602.20011
* Jump-diffusion generative model for financial time series that captures abrupt movements, heavy tails, and regime changes that diffusion-only models miss. Both drift and jump intensity learned from data. Applied to financial and energy series.
* TRADEABLE IDEA: Use the jump-detection component as a regime indicator. When the model's learned jump intensity spikes, it signals elevated probability of discontinuous moves -- widen stops or reduce size. Could be simplified: track realized jump frequency (moves > 2 sigma in < 1 min) as a rolling regime indicator for ES.
* Difficulty: Medium

5. Detecting Unlawful Insider Trading via Shapley Values and Causal Forests

* Link: https://arxiv.org/abs/2602.19841
* Uses SHAP + Causal Forests to identify features that predict informed trading. Key features: director status, price-to-book, recent returns, and market beta are statistically significant predictors of insider trading activity.
* TRADEABLE IDEA: Monitor SEC Form 4 filings for insider buys in stocks with high information asymmetry (low analyst

coverage, high P/B divergence, elevated beta). Cluster unusual insider buying -> go long those names or use as a sentiment overlay for sector rotation that feeds into ES directional bias.

* Difficulty: Easy