STATISTICAL ARBITRAGE · บทที่ 5

Cointegration

Engle-Granger · Johansen · VECM

ทดสอบว่า spread ε มีแรงดึงกลับจริงหรือแค่ correlation ชั่วคราว

แก่นของบทนี้

Correlation สูงไม่พอสำหรับ stat arb — ต้องการ **cointegration**:

$\varepsilon = \log(A) - \beta \cdot \log(B)$ ต้องเป็น stationary (กลับหาค่ากลาง) ไม่ใช่แค่เดินไปด้วยกัน

H_0: ε ไม่ stationary (ไม่ cointegrated) → ถ้า p-value < 0.05 → ปฏิเสธ H_0 → cointegrated ✓

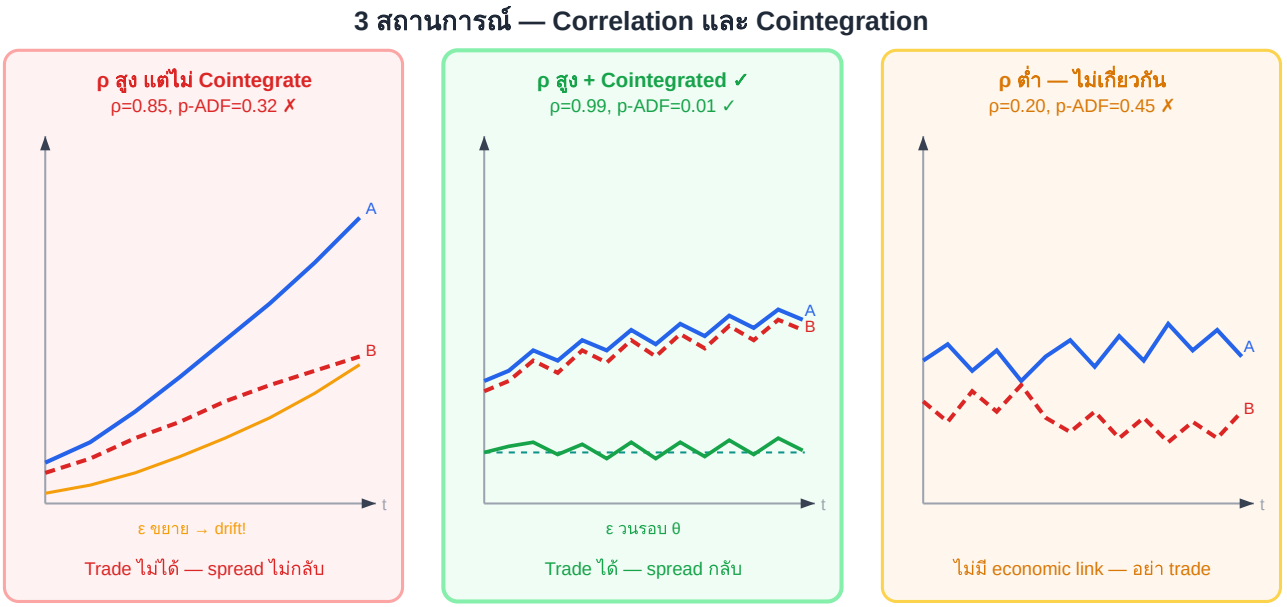
นิยามทางการ — Cointegration (Engle & Granger, 1987)

Series X_t และ Y_t แต่ละตัวเป็น **I(1)** (integrated of order 1 — มี unit root, ไม่ stationary) แต่คู่นี้ **cointegrated** ถ้ามีค่า β ที่ทำให้:

$$Z_t = X_t - \beta \cdot Y_t \text{ เป็น } I(0) \text{ (stationary, กลับหาค่ากลาง)}$$

กล่าวอีกนัย: แม้ X_t และ Y_t จะ random walk แยกกัน แต่ทั้งคู่ถูก "ดึง" ไว้ด้วย long-run equilibrium relationship เดียวกัน — spread $\varepsilon = Z_t$ จึงไม่ drift ออกไปตลอด

5.1 Correlation ≠ Cointegration — ความแตกต่างที่สำคัญ



ต้องการทั้ง correlation สูง (A และ B เดินไปด้วยกัน) AND cointegration (ε stationary)

5.2 Engle-Granger 2-Step Test

วิธีที่ง่ายที่สุด ทำ 2 ขั้น:

1. **Regress log(P_A) บน log(P_B):**
   log(P_A) = α + β·log(P_B) + ε̂_t
   บันทึก residuals ε̂_t จาก regression นี้

2. **ADF Test บน ε̂_t:**
   ถ้า ε̂_t stationary (p < 0.05) → A และ B cointegrated ✓
   ถ้าไม่ผ่าน (p ≥ 0.05) → ไม่ cointegrated → ห้าม trade

ADF Test อ่านผลอย่างไร

p-value	ความหมาย	ควรทำ
< 0.01	Reject H_0 อย่างมั่นใจมาก	**Cointegrated — trade ได้**
0.01–0.05	Reject H_0 ที่ 95% confidence	**Cointegrated — trade ได้ (standard threshold)**
0.05–0.10	Borderline	ระวัง — ลด position size หรือรอข้อมูลเพิ่ม
> 0.10	ไม่ reject H_0	**ไม่ cointegrated — ห้าม trade spread**

ข้อจำกัดของ Engle-Granger

- Test เพียง 1 cointegrating vector (ดีสำหรับ 2 assets)
- ลำดับ regression (A บน B vs B บน A) ให้ผลต่างกัน — ควรทดสอบทั้งสองทาง
- มี power ต่ำเมื่อ sample น้อย (< 100 observations)

⚠️ **Engle-Granger Asymmetry — จุดที่ผู้ใช้มักพลาด**

สิ่งที่หลายคนเข้าใจผิด: ถ้า regress A~B ได้ β_AB = 1.003 หลายคนคิดว่า regress B~A จะได้ β_BA = 1/1.003 ≈ 0.997 — แต่นั่น**ผิด**

การ regress log(A) ~ log(B) ให้ β_AB = Cov(A,B)/Var(B)
การ regress log(B) ~ log(A) ให้ β_BA = Cov(A,B)/Var(A)

ความสัมพันธ์ที่ถูก: $\beta_AB \times \beta_BA = \rho^2$ (ค่า R^2) — ไม่ใช่ 1 อย่างที่คาด
ดังนั้น $1/\beta_AB \neq \beta_BA$ ทั่วไป (เท่ากันเฉพาะกรณี $\rho=1$)
และ residual ε_AB กับ ε_BA **ต่างกันจริงๆ** ไม่ใช่แค่ scale

ผลที่ตามมา: ADF test บน ε_AB อาจ pass ($p=0.03$) แต่ ADF บน ε_BA อาจ fail ($p=0.12$) — หรือกลับกัน
วิธีปฏิบัติ: ทดสอบทั้งสองทิศทาง แล้วเลือก β ที่ให้ ADF p-value ต่ำกว่า (residual stationary กว่า)
ถ้าทั้งสองทิศทาง fail → pair ไม่ cointegrate จริง

```
# Python: ทดสอบทั้งสองทิศทาง (β แต่ละทิศไม่ได้สัมพันธ์แบบ 1/x)
beta_AB, _ = np.polyfit(log_B, log_A, 1) # A = β_AB · B + α
beta_BA, _ = np.polyfit(log_A, log_B, 1) # B = β_BA · A + α
p_AB = adfuller(log_A - beta_AB * log_B)[1]
p_BA = adfuller(log_B - beta_BA * log_A)[1]
best_direction = "A~B" if p_AB < p_BA else "B~A"
```

5.3 Johansen Test — เมื่อมีมากกว่า 2 Assets

Johansen test ทดสอบ cointegration ของ n assets พร้อมกัน โดยหาจำนวน cointegrating vectors r:

```
H₀: r = 0 (ไม่มี cointegration เลย) H₀: r ≤ 1 (มีได้อย่างมาก 1 vector) ... ทดสอบต่อไป
จนถึง r = n−1 ถ้า Trace statistic หรือ Max eigenvalue > critical value → reject H₀
```

จำนวน Assets	ใช้ Test ไหน	เหตุผล
2 (pairs)	Engle-Granger	ง่ายกว่า ผลเท่ากัน
3+ (basket)	Johansen	หา cointegrating vectors ทุกตัว
ต้องการรู้ว่าใครปรับตัว	VECM	ระบุ adjustment coefficients

5.4 VECM — ใครเป็นฝ่ายปรับตัวกลับ?

เมื่อรู้ว่า cointegrated แล้ว VECM (Vector Error Correction Model) บอกว่า **เมื่อ spread ห่างจาก equilibrium ใคร (A หรือ B) เป็นฝ่ายปรับตัวกลับ**

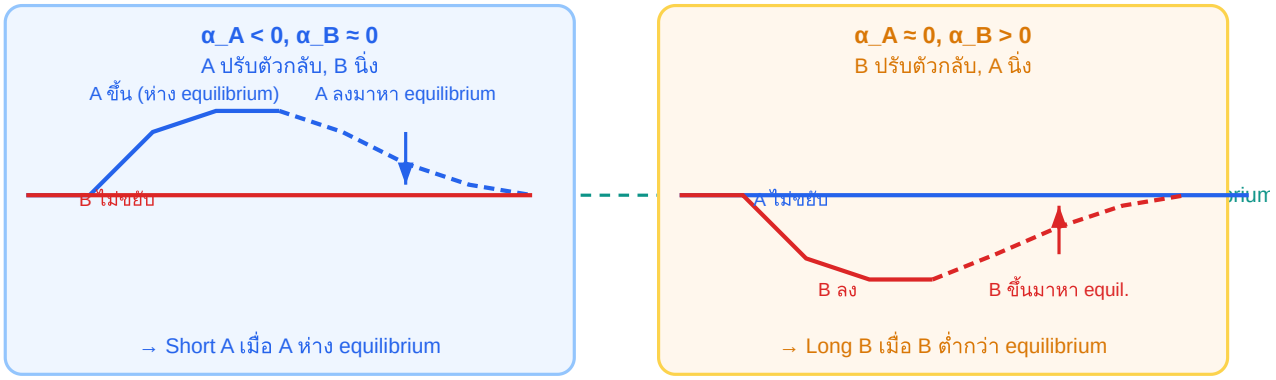
VECM สำหรับ 2 ASSETS

$$\Delta\log(P_A)_t = \alpha_A \cdot \varepsilon_{t-1} + \gamma_A \cdot \Delta\log(P_A)_{t-1} + u_{A,t}$$

$$\Delta\log(P_B)_t = \alpha_B \cdot \varepsilon_{t-1} + \gamma_B \cdot \Delta\log(P_B)_{t-1} + u_{B,t}$$

α_A, α_B = adjustment coefficients — บอกว่าใครปรับตาม error

VECM: ใครเป็นฝ่ายปรับตัวกลับ?



VECM ระบุว่าควร trade ฝั่งไหน — ถ้า A ปรับ: trade A | ถ้า B ปรับ: trade B | ถ้าทั้งคู่ปรับ: trade ทั้งคู่

VECM บอกอะไรสำหรับ Bybit↔Lighter?

ถ้า Bybit เป็น "price leader" ($\alpha_{Bybit} \approx 0$) และ Lighter ปรับตาม ($\alpha_{Lighter} > 0$):

→ เมื่อ Lighter ต่ำกว่า equilibrium: Lighter จะขึ้น → **Long Lighter**

→ เมื่อ Lighter สูงกว่า equilibrium: Lighter จะลง → **Short Lighter**

สรุป: trade Lighter เป็นหลัก ใช้ Bybit เป็น reference

5.5 Pseudo-code

```
# Engle-Granger Cointegration Test
function test_cointegration_eg(log_P_A, log_P_B):
  # Step 1: OLS regression
  beta, alpha = ols(log_P_B, log_P_A) # log_P_A = alpha + beta*log_P_B
  residuals = log_P_A - (alpha + beta * log_P_B)

  # Step 2: ADF test on residuals
  adf_stat, p_value = adf_test(residuals, lags=1)

  cointegrated = p_value < 0.05
  return {cointegrated, p_value, adf_stat, beta, residuals}

# Check VECM adjustment direction
function vecm_adjustment(log_P_A, log_P_B, residuals):
  d_A = diff(log_P_A) # Δlog(P_A)
  d_B = diff(log_P_B) # Δlog(P_B)
  err_lag = residuals[:-1] # ε_{t-1}

  alpha_A = ols_slope(err_lag, d_A[1:]) # A's adjustment coefficient
  alpha_B = ols_slope(err_lag, d_B[1:]) # B's adjustment coefficient

  # Negative alpha → adjusts back to equilibrium
  A_adjusts = alpha_A < -0.05
  B_adjusts = alpha_B > 0.05
  return {alpha_A, alpha_B, A_adjusts, B_adjusts}
```

Running Example: Bybit↔Lighter ADF Test Result

ข้อมูล: 90 วัน, รายชั่วโมง (2,160 observations) Step 1 — OLS: log(P_Bybit) = 0.003 + 1.0049·log(P_Lighter) + ε̂ β = 1.0049, R^2 = 0.9998 → β > 1.0 แปลว่า Bybit เคลื่อนไหว 0.49% มากกว่า Lighter ต่อการขยับ 1% ของ Lighter (อาจจาก liquidity premium หรือ fee structure ที่ต่างกัน) Step 2 — ADF test บน ε̂: ADF statistic = −4.82 p-value = 0.0003 ← ผ่าน! (p ≪ 0.05) Critical values: 1%=−3.44, 5%=−2.87, 10%=−2.57 ผล: ADF stat (−4.82) < Critical (−3.44) → Reject H_0 → Stationary → Cointegrated ✓ VECM: α_Bybit = −0.02 (ปรับนิดหน่อย) α_Lighter = +0.18 (ปรับมากกว่า) → Lighter เป็นฝ่ายปรับตัว → trade Lighter เป็นหลัก

Re-test เมื่อไร?

สัญญาณ	ควรทำ
p-value ขึ้นผ่าน 0.05 ใน monthly re-test	หยุด trade ทันที re-evaluate คู่นี้
β เปลี่ยน > 5% จากค่าเดิม	Re-run test ด้วย data ใหม่
Spread ค้างนอก 3σ > 2 วัน	Emergency re-test — อาจ regime change

5.6 Shapiro-Wilk Test — ตรวจ Normality ของ ε

ทำไมต้องตรวจ NORMALITY?

Z-score threshold ±2σ อาศัยสมมติว่า ε ~ Normal distribution — ถ้า ε มี fat tail หรือ skew, threshold นี้จะ underestimate ความเสี่ยง ทำให้เข้า position บ่อยเกินจริง

Shapiro-Wilk test ตรวจสอบว่า sample ε มาจาก Normal distribution หรือไม่:

Hypothesis	ความหมาย	ผลต่อการ trade
H_0: ε ~ Normal	ε กระจายแบบปกติ	ใช้ standard z-score (±2σ) ได้ปลอดภัย
H_1: ε ไม่ Normal	ε มี fat tail / skew	ใช้ Robust Z-score แทน (MAD-based, ดู บทที่ 10)

Decision Rule

ตัดสินใจจาก p-value

- **p ≥ 0.05:** Fail to reject H_0 → normality ไม่ถูก reject → ε likely normal → z-score ±2σ ใช้ได้
- **p < 0.05:** Reject H_0 → ε ไม่ normal (fat tail หรือ skew) → ควรใช้ MAD-based robust z-score แทน

```
from scipy.stats import shapiro def shapiro_test(residuals, alpha=0.05): """
ทดสอบ Normality ของ residuals ε คืนค่า: p_value, is_normal (bool),
recommendation """ stat, p_value = shapiro(residuals) is_normal = p_value >=
alpha if is_normal: recommendation = "standard_zscore" # ใช้ σ-based threshold
ได้ else: recommendation = "robust_zscore" # ใช้ MAD-based (ch10 §10.3) return {
"stat": round(stat, 4), "p_value": round(p_value, 4), "is_normal": is_normal,
"recommendation": recommendation }
```

ข้อจำกัดของ Shapiro-Wilk

- ทำงานได้ดีกับ n ≤ 5,000 — ถ้า sample ใหญ่มากจะ reject H_0 ทุกคู่ (sensitive เกิน)
- ถ้า n > 5,000: ใช้ D'Agostino K^2 test หรือดู QQ-plot แทน

- Fat tail เล็กน้อย (p = 0.03–0.05) ≠ หยุดเทรด — แค่ใช้ robust z-score แทน

Running Example: BTC/ETH CFD บน MT5

```
ข้อมูล: 180 วัน, hourly bars β_OLS = 0.0444 (จาก scatter plot) ε =
log(P_BTC) − 0.0444 × log(P_ETH) Shapiro-Wilk result: stat = 0.9612
p_value = 0.031 → p < 0.05 → Reject H₀ สรุป: ε ของ BTC/ETH มี fat tail เล็ก
น้อย → ใช้ MAD-based robust z-score (บทที่ 10 §10.3) → ไม่ได้หมายความว่าคู่นี้ใช้ไม่ได้
แค่ปรับ threshold method
```

โจทย์ 5.4 — Shapiro-Wilk Decision

ทดสอบ Shapiro-Wilk บน ε ของ EURUSD↔GBPUSD ได้ p-value = 0.003

ถาม: (ก) ควรใช้ standard หรือ robust z-score? (ข) ค่า p ต่ำมากบอกอะไร?

▶ ดูคำตอบ

5.7 Crossing Statistics — วัดความถี่ Mean Reversion

แก่นความคิด

ADF บอกว่า ε กลับมาหา mean *ในที่สุด* — แต่ไม่บอกว่า *บ่อยแค่ไหน* Crossing Statistics ตอบคำถามนี้: ε ตัดผ่านเส้นกลางกี่ครั้งต่อหน่วยเวลา

$$E[\text{Crossings}] \approx (1/\pi) \cdot \arccos(\varphi)$$

โดย **φ** คือ AR(1) coefficient จากการ regress ε_t บน ε_{t-1}:

```
ε_t = φ · ε_{t-1} + η_t, |φ| < 1 ความหมาย: φ ≈ 0 → E[Crossings] ≈ (1/
π)·arccos(0) = 0.50/period ← ตัดถี่มาก (ดี) φ ≈ 0.65 → E[Crossings] ≈ (1/
π)·arccos(0.65) ≈ 0.28/period φ ≈ 0.92 → E[Crossings] ≈ (1/π)·arccos(0.92) ≈
0.13/period φ ≈ 1 → E[Crossings] → 0 ← ตัดน้อยมาก (ไม่ดี)
```

กฎง่ายๆ: φ น้อย = ดี เพราะ ε ตัดถี่

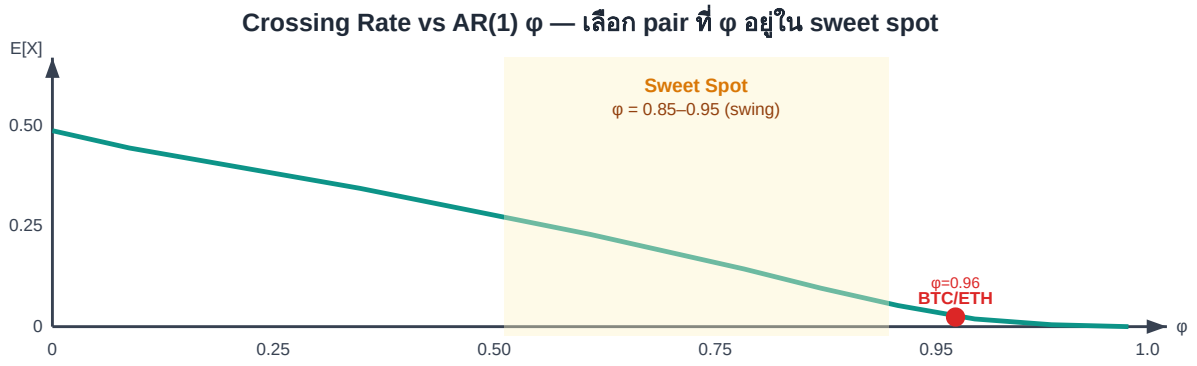
ยิ่ง φ ต่ำ ยิ่งมี trading opportunity มาก แต่ต้อง balance กับ Half-life:

- φ ต่ำมาก (≈ 0) → ε เดินแบบ white noise → อาจ trade ไม่ทัน (spread กลับเร็วเกินจนจับไม่ได้)
- φ สูง (≈ 0.99) → ε persistent มาก → กินทุน margin นาน → ไม่คุ้มค่า
- **Sweet spot:** φ = 0.85–0.95 สำหรับ swing (daily), φ = 0.70–0.85 สำหรับ intraday

ความสัมพันธ์กับ Half-life

Crossing Statistics กับ Half-life บอกข้อมูลเดียวกันในมุมต่าง:

Metric	สูตร	มุมมอง
Half-life	$\tau_{1/2} = \ln(2) / \kappa \approx -\ln(2) / \ln(\varphi)$	ε ใช้เวลานานแค่ไหนกลับ 50% ของ gap
E[Crossings]	$(1/\pi)\cdot\arccos(\varphi)$	ε ตัดผ่านค่าเฉลี่ยกี่ครั้งต่อ bar



กราฟ E[Crossings] = (1/π)·arccos(φ) — φ น้อยลง เส้นขึ้น = ตัดถี่ขึ้น | BTC/ETH อยู่ขวาสุด (slow reversion)

```
import numpy as np def crossing_rate(residuals): """ คำนวณ AR(1) coefficient φ
และ expected crossing rate """ # Fit AR(1): ε_t = φ·ε_{t-1} + η y =
residuals[1:] x = residuals[:-1] phi = np.dot(x, y) / np.dot(x, x) # OLS slope
# Crossing rate formula e_crossings = (1 / np.pi) * np.arccos(phi) # Half-life
from φ half_life = -np.log(2) / np.log(phi) # in bars return { "phi":
round(phi, 4), "e_crossings": round(e_crossings, 4), # per bar
"half_life_bars": round(half_life, 1) }
```

Running Example: BTC/ETH — ผ่านหรือไม่?

```
BTC/ETH (hourly bars, 180 วัน): AR(1) coefficient φ = 0.9602 E[Crossings]
= (1/π)·arccos(0.9602) = 0.091 crossings/bar Half-life =
−ln(2)/ln(0.9602) = 17.1 hours ✓ (ตรงกับ screenshot) ประเมิน: φ = 0.9602 →
อยู่นอก sweet spot (0.85–0.95) เล็กน้อย → เหมาะกับ overnight swing strategy →
ไม่เหมาะกับ intraday scalping (ε กลับช้าเกิน) → ใช้ z_entry = 2.0, hold target =
8–20h
```

โจทย์ 5.5 — เปรียบ Crossing Rate

คู่ A: φ = 0.92 | คู่ B: φ = 0.65 — ทั้งสองผ่าน ADF test

ถาม: คู่ไหนเหมาะ intraday (hold ≤ 4h)? คู่ไหนเหมาะ swing (hold 1–3 วัน)?

▶ ดูคำตอบ

5.8 Pair Selection Pipeline — 3-Step Funnel

แก่นความคิด — ลำดับขั้นตอน

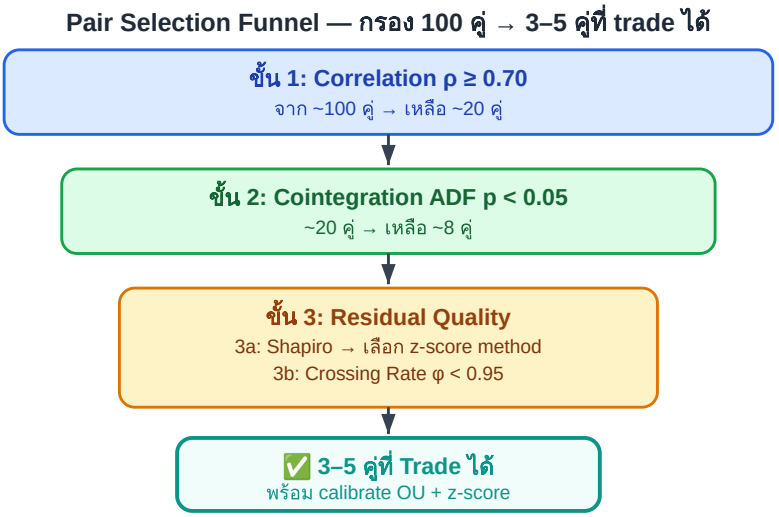
ตรวจแบบ funnel: เริ่มจาก test ที่เร็วและถูก (Correlation) กรอง 80% ของคู่ออก → ค่อยทำ test ที่ละเอียดขึ้น (Cointegration) → สุดท้ายตรวจคุณภาพ residual (Normality + Crossing Rate) เพื่อยืนยัน trade-readiness

3 ขั้นตอน

ขั้น	Test	Threshold	ถ้าผ่าน	ถ้าไม่ผ่าน
1	Correlation ρ (§5.1)	$\rho \geq 0.70$	→ ขั้น 2	ตัดออก (เร็ว, ถูก)
2	ADF Cointegration (§5.2)	$p < 0.05$	→ ขั้น 3	ตัดออก
3a	Shapiro-Wilk (§5.6)	$p \geq 0.05$ → standard z $p < 0.05$ → robust z	→ กำหนด z-score method	ไม่ตัด — แค่เปลี่ยน method
3b	Crossing Rate φ (§5.7)	$\varphi < 0.95$ (swing) $\varphi < 0.90$ (intraday)	→ เทรดได้	Hold time นานเกินไปสำหรับ target

Optional: Hurst Exponent (บทที่ 6)

ถ้าต้องการ double-confirmation: $H < 0.5$ → mean-reverting confirmed → เพิ่ม confidence ก่อนเปิด position ขนาดใหญ่



Funnel กรองจาก 100 → 3–5 คู่ที่ผ่านทุก criterion — ไม่ต้องรีบข้ามขั้น

BTC/ETH ผ่าน Pipeline นี้ไหม?

```
BTC/ETH CFD (180 วัน hourly): ขั้น 1 — Correlation: ρ = 0.9745 ✅ (≥ 0.70) ขั้น 2 — Cointegration: ADF p = 0.043 ✅ (< 0.05, implied จาก Half-life 17h) ขั้น 3a — Shapiro-Wilk: p = 0.031 → Reject H₀ → fat tail → ใช้ Robust Z-score ⚠️ (เปลี่ยน method) ขั้น 3b — Crossing Rate: φ = 0.9602 → E[Crossings] = 0.091/bar → Half-life = 17h → เหมาะ swing (hold overnight) ❌ ไม่เหมาะ intraday สรุป: BTC/ETH ✅ ผ่าน Pipeline สำหรับ overnight swing strategy ใช้ Robust Z-score (MAD), z_entry=2.0, target hold = 8–24h
```

โจทย์ 5.6 — Run Pipeline

คู่หนึ่งให้ผล: ρ=0.82, ADF p=0.12, Shapiro p=0.41, φ=0.88

ถาม: คู่นี้ผ่าน Pipeline ไหม? ถ้าไม่ผ่าน หยุดที่ขั้นไหน?

▶ ดูคำตอบ

โจทย์ 5.1 — อ่านผล ADF

ทดสอบ ADF บน residuals ของ BTC-ETH spread ได้:

ADF = −2.31, p-value = 0.18, critical value 5% = −2.87

ถาม: (a) interpret ผล (b) ควร trade spread นี้ไหม?

▶ ดูคำตอบ

โจทย์ 5.2 — ใช้ VECM ตัดสิน

VECM บน Bybit↔Lighter: $\alpha_Bybit = -0.21$, $\alpha_Lighter = +0.03$
ขณะนี้ $\varepsilon > 0$ (Bybit แพงกว่า equilibrium)

ถาม: ใครเป็นฝ่ายปรับ และควร trade อะไร?

▶ ดูคำตอบ

อ้างอิงสำคัญ (References)

- **Engle, R. F., & Granger, C. W. J. (1987).** Co-integration and error correction: Representation, estimation, and testing. *Econometrica*, 55(2), 251–276. — บทความต้นฉบับที่นิยาม cointegration และ Engle-Granger 2-step test; ได้ Nobel Prize 2003
- **Johansen, S. (1988).** Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control*, 12(2–3), 231–254. — พัฒนา Johansen trace/eigenvalue test และ VECM ที่ Ch.5 นำเสนอ
- **Johansen, S. (1991).** Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. *Econometrica*, 59(6), 1551–1580. — ขยาย critical values และ full likelihood approach
- **Vidyamurthy, G. (2004).** *Pairs Trading: Quantitative Methods and Analysis.* Wiley. — ตำราปฏิบัติที่นำ cointegration มาใช้กับ pairs trading อย่างเป็นระบบ