Time-Series Analysis for Pairs Trading and Cointegration

trading the gap between two prices, not the direction of either

Martin Burke (with Claude)

Draft

Abstract

Most ways of making money in markets are a bet on direction: buy something you think will go up. This note is about a different kind of bet—one that is deliberately indifferent to whether the market rises or falls, and pays off instead when the *gap* between two related prices closes. The catch is that "related" has to mean something much stronger than "they look like they move together". Two prices can rise and fall in near-lockstep for years and still drift apart forever; betting on their gap is then a slow way to lose money. The property that actually licenses the trade is *cointegration*: although each price wanders unpredictably on its own, some fixed combination of the two is tethered to a constant level and keeps returning to it. This note builds the idea from the ground up. We start with why a single price is essentially untradeable on its own (it is a *random walk*), explain the statistical test that tells a genuine tether from a mirage, construct the trade from the tethered combination, and then scale the whole apparatus up from a pair to a whole basket of assets using vector autoregressions, the Johansen procedure and the vector error-correction model. Throughout, we keep an eye on the real failure mode—tethers that hold in the data you fit and snap in the data you trade.

Contents

1 The trade that doesn't care which way the market goes

Suppose you are convinced that Goldman Sachs and Morgan Stanley—two large, similar investment banks—are "worth about the same" in some loose sense, and you notice that on a given day Goldman looks expensive relative to Morgan. You could try to profit from that without taking any view at all on whether bank stocks, or the market as a whole, are about to go up or down: *sell* the expensive one and *buy* the cheap one, in the right proportion, and wait for the discrepancy to close. If the whole sector rallies, your long leg gains and your short leg loses, and the two roughly cancel; if the sector sells off, the same cancellation works in reverse. What you are left exposed to is only the *relative* mispricing—the gap—and your bet is simply that the gap will shrink back to normal. This is **pairs trading**, the canonical example of a *market-neutral* or *statistical-arbitrage* strategy, and it is the thread that will pull the entire toolbox of time-series analysis into view.

Figure 1 shows why the idea is tempting. The top panel is the daily price of Goldman and Morgan over five years, each rebased to start at 1.0. They move together closely—the same rallies, the same sell-offs. The bottom panel is the quantity that matters for the trade: a particular combination of the two, $s_t = \mathrm{GS}_t - 3.30\,\mathrm{MS}_t$, which we will call the *spread*. Notice what the spread does. It does not wander off; it oscillates around a fixed level, drifting up and down but always being hauled back. That return-to-a-level behaviour is the whole game. When the spread is unusually high, you sell it (short Goldman, buy Morgan); when it is unusually low, you buy it; and you collect the difference when it reverts. The trade is an automated bet on the spread's *mean reversion*.

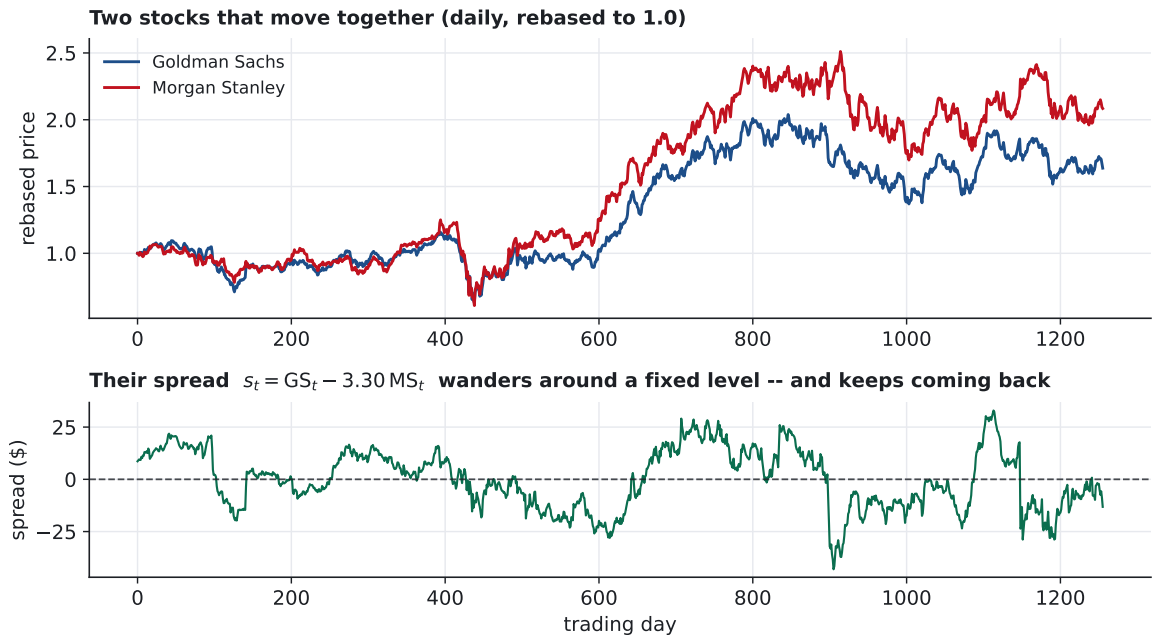

Figure 1: The promise of pairs trading. **Top:** two similar bank stocks move together. **Bottom:** a fixed combination of them—the spread $s_t = \mathrm{GS}_t - 3.30\,\mathrm{MS}_t$—does not wander off but oscillates around a constant level and keeps coming back to it. A strategy that sells the spread when it is high and buys it when it is low is betting only on that reversion, not on the direction of the market. Everything in this note is about when that bottom panel is trustworthy—and when it is a mirage.

The rest of this note is an answer to one question hiding inside that picture: *when is the bottom panel real?* It is easy to draw two lines that look like they move together and subtract one from the other; the dangerous part is that the result can look mean-reverting over the sample you happened to look at and yet have no tendency to revert at all going forward. Distinguishing a genuine, tradeable tether from a coincidence is not a matter of eyeballing charts—it requires

the machinery of time-series analysis, and in particular the idea of *cointegration*. Building that idea carefully, and showing how to test for it, is the core of what follows.

The plan. Part I works entirely with a single pair, because every essential idea is already visible there. We first confront the trap that the most obvious measure of "moving together"—correlation—is the wrong one (Section 2). We then explain why an individual price is untradeable on its own, because it is a *random walk* with no level to revert to (Section 3), and meet the test that detects this (Section 4). Cointegration is the precise statement that two such untradeable series can combine into a tradeable one (Section 5), and the *spread* we trade is exactly that combination; we turn it into a signal and a trade in Sections 6–7. Part II then scales up from two assets to many—vector autoregressions, lead–lag structure, and the Johansen machinery that finds several tethers at once in a basket (Sections 8–11). Part III is the cautionary part: tethers break, and the section on it is load-bearing, not a disclaimer (Section 12). Appendices hold the statistical machinery—the unusual distribution behind the unit-root test, the spurious-regression phenomenon, and the algebra of the error-correction model—so the main line can stay about the ideas.

Who this is for. You are comfortable with basic probability, ordinary least-squares regression and a little linear algebra. You are *not* assumed to know any finance beyond what a share price is, nor any specialist time-series theory—every term of art (stationarity, unit root, cointegration, error correction) is built up from scratch. No stochastic calculus is used anywhere.

Part I — One pair

2 The trap: correlation is not cointegration

The intuitive way to say two stocks "move together" is that they are *correlated*: when one has a good day, so does the other. Correlation is computed on the day-to-day *changes* (returns), and it is the first thing anyone reaches for. It is also, for our purpose, a trap. Correlation measures whether two series move in the same direction *on the same day*; it says nothing about whether they stay close *over time*. Those are completely different questions, and conflating them is the single most common way to build a pairs strategy that quietly bleeds.

Figure 2 makes the distinction concrete with two simulated pairs. The left pair has strongly correlated daily moves—a return correlation near 0.5, the kind of number that would pass a casual screen—and yet the two prices pull apart and their difference wanders away, never settling. There is no level for a spread trade to revert to; a trader who shorted the gap early would have watched it widen indefinitely. The right pair is different in a way that correlation cannot see: the two prices are tied together so that a fixed combination of them is pinned to a constant level. *That* difference reverts. The two panels can have identical correlations; what separates them is whether a combination is tethered, and that is the property we need.

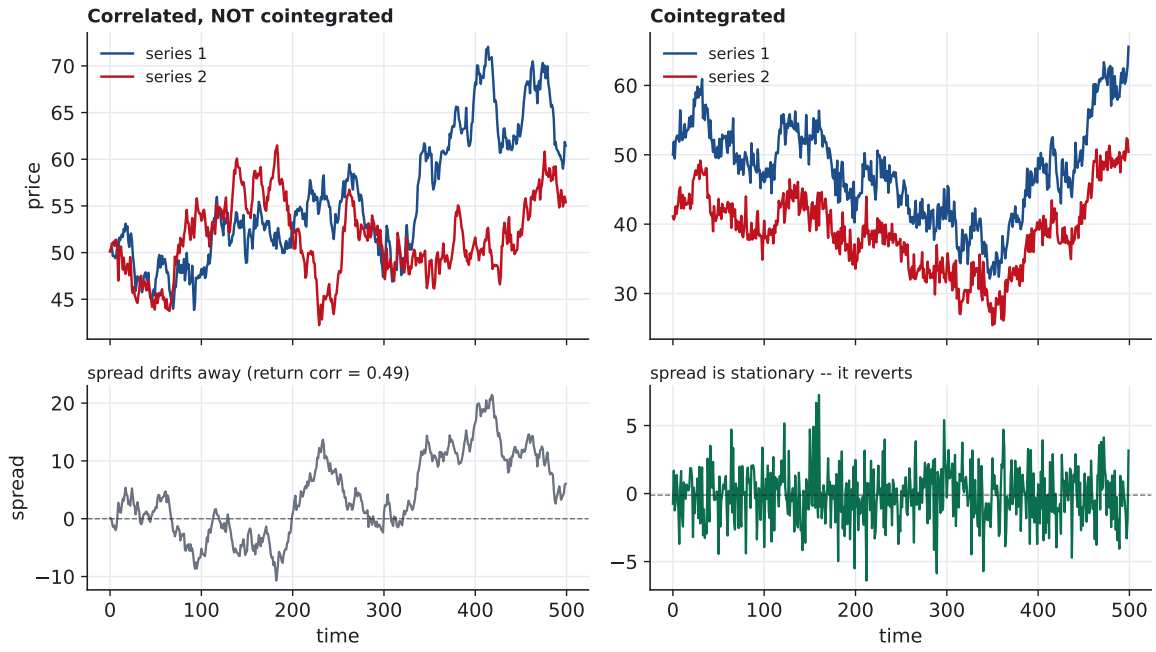

Figure 2: Why correlation is the wrong test. **Left:** two series whose daily moves are strongly correlated, yet whose spread drifts away and never comes back—*not* tradeable. **Right:** two series tied by a fixed combination that stays pinned to a level—their spread reverts, and *is* tradeable. Correlation looks at same-day co-movement; cointegration looks at whether the two stay tethered over time. They are different properties, and only the second one licenses a pairs trade.

The hinge of the whole note. *Correlation* is about short-term, same-day co-movement of returns. *Cointegration* is about a long-term tether: a fixed combination of the prices that refuses to wander. A pairs trade lives or dies on the second property, not the first. Two assets can be highly correlated and not cointegrated (left panel above), or cointegrated without being especially correlated day to day. We spend the next three sections making "refuses to wander" precise.

To make "refuses to wander" precise, we need two technical ideas: what it means for a single series to wander (a *unit root*), and what it means for a combination not to (*stationarity*). We take them in turn.

3 Why a single price wanders: random walks and unit roots

In plain terms. A share price has no gravitational pull toward any particular value. Today's price is the best forecast of tomorrow's; the change from one day to the next is essentially unpredictable noise. A series built this way—next value equals current value plus fresh noise—is a *random walk*. It has no fixed level it tends to return to, so on its own it is useless for a mean-reversion trade: there is nothing to revert to. The formal name for the property that makes it wander is a *unit root*.

A random walk is the simplest model of a price:

$$y_t = y_{t-1} + \varepsilon_t, \tag{1}$$

where ε_t is a fresh, zero-mean shock each period, independent of the past. Equation (1) is a special case of the first-order autoregression (an AR(1) process)

$$y_t = \phi\, y_{t-1} + \varepsilon_t, \tag{2}$$

in which each value is a fraction ϕ of the previous one plus noise. The single number ϕ decides the entire character of the series. If $|\phi| < 1$, any displacement is damped: a value above zero tends to be followed by a smaller one, and the series is pulled back toward zero. It has a stable mean and a stable variance; we call such a series *stationary*. If $\phi = 1$ exactly—the random-walk case—the shocks never fade. Each ε_t is baked into the level permanently, the displacements accumulate, and the series wanders with ever growing range. That boundary value, $\phi = 1$, is the **unit root**, and it is the mathematical signature of a series that does not revert.

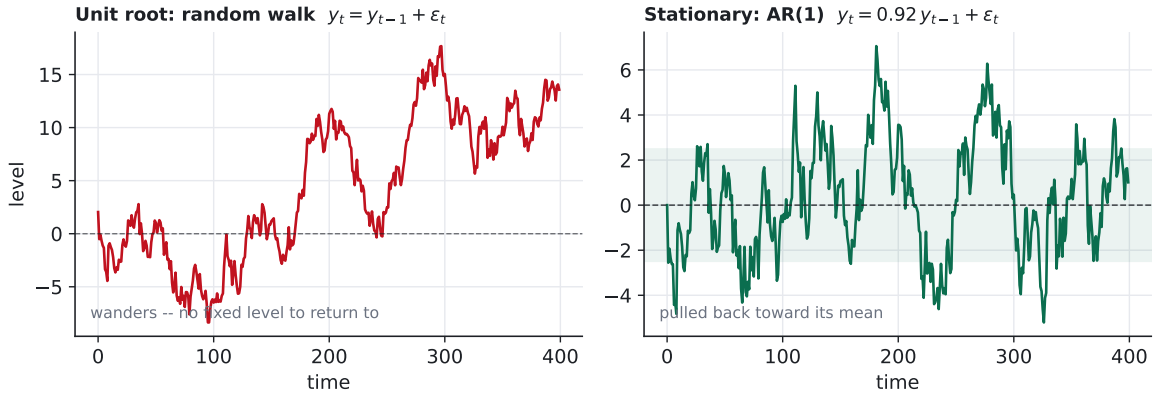

Figure 3: The same noise, two regimes. **Left:** a unit-root random walk ($\phi = 1$); shocks accumulate, and the series wanders with no level to come back to. **Right:** a stationary AR(1) ($\phi = 0.92$); shocks fade, and the series is repeatedly pulled back toward its mean (shaded band: ± 1 standard deviation). A single asset price behaves like the left panel; the spread of a good pair behaves like the right.

Figure 3 drives the same shocks through both regimes. With $\phi = 1$ the path roams; with $\phi = 0.92$ it oscillates inside a band around its mean. The visual difference is the whole distinction between an untradeable level and a tradeable one. Real asset prices look overwhelmingly like the left panel: to a very good approximation a price is a random walk, which is just a restatement of the old observation that you cannot predict tomorrow's return from the chart.

Differencing, and the language of integration. There is one reliable way to turn a wandering series into a well-behaved one: look at its *changes* rather than its levels. If y_t is a random walk, then its first difference $\Delta y_t = y_t - y_{t-1} = \varepsilon_t$ is just the noise— stationary by construction. A series that is non-stationary in levels but becomes stationary after differencing once is called *integrated of order one*, written $I(1)$; a series that is already stationary is $I(0)$. Prices are typically $I(1)$ (you must difference once, into returns, to get something stable); returns are $I(0)$. This vocabulary—$I(1)$ for "wanders, needs one difference", $I(0)$ for "already stable"—is exactly what we need to state cointegration crisply in Section 5.

4 Detecting a unit root: the Augmented Dickey–Fuller test

In plain terms. We need a statistical test that, handed a series, tells us whether it has a unit root (wanders, $I(1)$) or not (reverts, $I(0)$). The idea is simple: estimate something like ϕ from the data and check whether it is convincingly below 1. The execution has one famous subtlety—the usual "is this coefficient significant?" table from ordinary regression does *not* apply here—which we flag now and explain fully in Appendix A.

Rearranging the autoregression (2) by subtracting y_{t-1} from both sides turns the question "is $\phi = 1$?" into "is the coefficient on y_{t-1} zero?", which is easier to test. Writing $\rho = \phi - 1$,

$$\Delta y_t = \rho\, y_{t-1} + \varepsilon_t, \qquad \rho = \phi - 1. \tag{3}$$

A unit root ($\phi = 1$) is now the hypothesis $\rho = 0$, and a reverting series is $\rho < 0$ (a level above the mean produces a negative expected change, pulling it back). The **Dickey–Fuller test** estimates ρ by ordinary least squares and forms the t-ratio $\hat{\rho}/\text{s.e.}(\hat{\rho})$. A strongly negative value is evidence *against* the unit root—evidence that the series reverts. The *Augmented* Dickey–Fuller (ADF) test adds a few lagged differences Δy_{t-k} to soak up short-term autocorrelation, and usually a constant and possibly a time trend, so that the bare test for the unit root is not contaminated by those nuisance features:

$$\Delta y_t = C + \delta\, t + \rho\, y_{t-1} + \sum_{k=1}^{p} \beta_k\, \Delta y_{t-k} + \varepsilon_t. \tag{4}$$

The object of interest is still the single coefficient ρ on the lagged level, and the test statistic is still its t-ratio.

The catch (deferred to an appendix). Under the null hypothesis of a unit root, that t-ratio does *not* follow the bell-shaped distribution you would use to judge significance in an ordinary regression. It follows a different, left-shifted distribution—the *Dickey–Fuller distribution*—so the familiar cut-off of -1.96 is wrong and far too lenient. The correct 5% cut-offs are substantially more negative (around -2.86 with a constant, or -3.41 once a trend is included). Why this happens—and why ignoring it is a recipe for "finding" tradeable relationships that aren't there—is the subject of Appendix A; for now we simply use the right critical values.

The unit-root test, in one line. Regress the change Δy_t on the lagged level y_{t-1} (plus housekeeping terms) and look at the t-ratio of the lagged-level coefficient. Very negative $\Rightarrow$ the series reverts ($I(0)$, stationary). Near zero $\Rightarrow$ it wanders ($I(1)$, a unit root). Compare against the special Dickey–Fuller critical values, never the ordinary ± 1.96.

We now have everything needed to define cointegration: a test that classifies any series as wandering or reverting, and the language ($I(1)$, $I(0)$) to talk about combinations of them.

5 Cointegration: a stationary mix of wandering series

In plain terms. Here is the small miracle that makes pairs trading possible. Take two series that each individually wander like a random walk—each $I(1)$, each with no level to revert to. In general any combination of them also wanders. But for special pairs there exists one particular combination—one specific hedge ratio—that is stationary: it reverts to a fixed level even though its two ingredients do not. When such a combination exists, the two series are **cointegrated**, and that stationary combination is precisely the spread we trade.

Formally, two $I(1)$ series p_{1t} and p_{2t} are cointegrated if there is a coefficient β such that

$$s_t = p_{1t} - \beta\, p_{2t} \quad \text{is stationary } (I(0)). \tag{5}$$

The vector $(1, -\beta)$ is the *cointegrating vector* and β is the *hedge ratio*: it says how many units of asset 2 to short against one unit of asset 1 so that the common wandering cancels out, leaving only the stationary, mean-reverting residual. Intuitively, both prices are driven by the same underlying stochastic trend (the fortunes of the banking sector, say); β is tuned so that subtracting $\beta\, p_{2t}$ removes that shared trend exactly, exposing the stationary gap that remains. This is the precise content of the right-hand panel of Figure 2, and the reason the bottom panel of Figure 1 reverts.

5.1 Finding the spread: the Engle–Granger two-step

Engle and Granger gave a disarmingly simple recipe for both estimating the hedge ratio and testing whether a pair is cointegrated at all [1]. It is two ordinary-regression steps, and it remains the workhorse for a single pair.

Step 1: estimate the hedge ratio. Regress one price on the other by ordinary least squares,

$$p_{1t} = \alpha + \beta\, p_{2t} + e_t. \tag{6}$$

The slope $\hat{\beta}$ is the hedge ratio; the fitted residual $\hat{e}_t = p_{1t} - \hat{\alpha} - \hat{\beta}\, p_{2t}$ is our candidate spread. The left panel of Figure 4 shows this for Goldman against Morgan: the cloud of daily points is tight around a line of slope $\hat{\beta} = 3.30$, meaning one share of Goldman is hedged by shorting about 3.3 shares of Morgan.

Step 2: test the residual for a unit root. Run the ADF test of Section 4 on the residual $\hat{e}_t$. If the residual is stationary, the two prices are cointegrated and the spread is tradeable; if the residual itself wanders, the apparent relationship is a mirage and there is no trade. For Goldman/Morgan the residual's ADF statistic is -3.50, below the relevant 5% critical value of -3.34,[1] so the pair passes: it is cointegrated, and the right-hand panel of Figure 4 shows the resulting stationary spread.

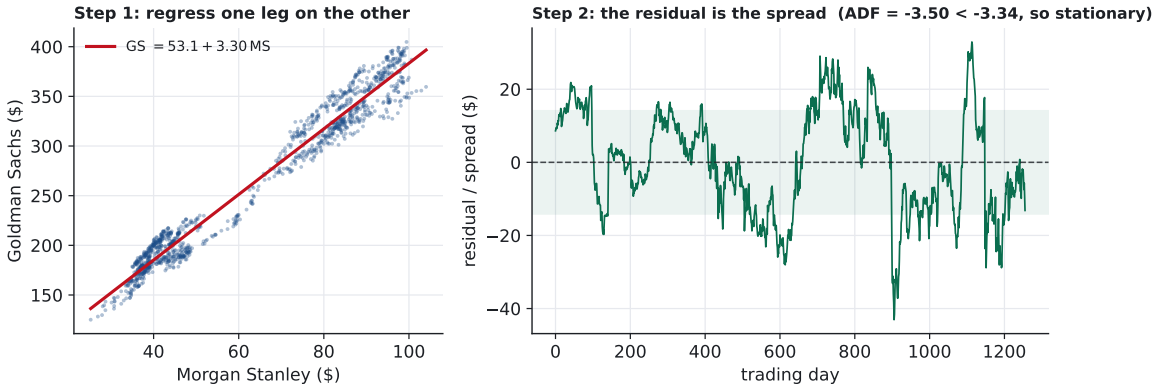

Figure 4: The Engle–Granger two-step on Goldman Sachs and Morgan Stanley. **Left, Step 1:** regress one leg on the other; the slope is the hedge ratio $\hat{\beta} = 3.30$. **Right, Step 2:** the regression residual is the candidate spread, and an ADF test on it returns $-3.50 < -3.34$, so it is stationary and the pair is cointegrated. The shaded band is ± 1 standard deviation of the spread—the natural scale for the trading signal in Section 7.

Engle–Granger in two lines. (1) OLS-regress p_1 on p_2; the slope is the hedge ratio β, the residual is the candidate spread. (2) ADF-test that residual: stationary $\Rightarrow$ cointegrated $\Rightarrow$ tradeable. Skip step 2 and you are back in the trap of Section 2—a regression of one wandering series on another almost always *looks* significant even when the two are unrelated (the spurious-regression problem, Appendix B).

6 From spread to signal: mean reversion and its speed

We have a stationary spread. To trade it we need to quantify two things: how far it currently sits from its normal level, and how quickly it tends to close that gap.

The z-score: how far is "far"? The natural yardstick is the spread's own variability. Subtract the spread's mean and divide by its standard deviation to get the *z-score*

$$z_t = \frac{s_t - \mu}{\sigma}, \tag{7}$$

[1] The critical value for an ADF test on an *estimated* residual is more negative than the ordinary Dickey–Fuller value, because Step 1 has already chosen β to make the residual look as stationary as possible—an advantage the test must correct for. These are the Engle–Granger critical values; see Appendix A.

which measures the dislocation in standard deviations. A z_t of $+2$ means the spread is two standard deviations rich; -2 means two cheap; 0 means it is at its normal level. The z-score is the trading signal, and because it is dimensionless it transfers unchanged from one pair to another. (In live trading, μ and σ are computed over a trailing window rather than the whole sample, so the yardstick adapts as the spread's level slowly evolves; we return to this in Section 7.) The left panel of Figure 5 shows the Goldman/Morgan spread expressed as a z-score, oscillating through ± 2.

The half-life: how quickly does it revert? Reversion is governed by the same ρ from the unit-root regression. Fitting $\Delta s_t = \rho\,(s_{t-1} - \mu)$ to the spread, a value $\rho < 0$ means a deviation of size d shrinks in expectation to $(1+\rho)\,d$ next period, and to $(1+\rho)^k d$ after k periods—geometric decay. The convenient summary is the *half-life*, the time for a deviation to halve:

$$\text{half-life} = \frac{\ln 2}{-\ln(1+\rho)} \approx \frac{\ln 2}{\kappa}, \qquad \kappa \equiv -\rho. \tag{8}$$

For Goldman/Morgan the fitted speed gives a half-life of about 38 trading days (right panel of Figure 5)—a deviation typically takes roughly two months to half-close. The half-life is the single most useful number for sizing a pairs trade: it sets the holding period, tells you how long your capital is committed per round trip, and warns you when a "cointegrated" spread is so sluggish that transaction costs and the risk of a structural break (Section 12) will eat the edge before reversion arrives.

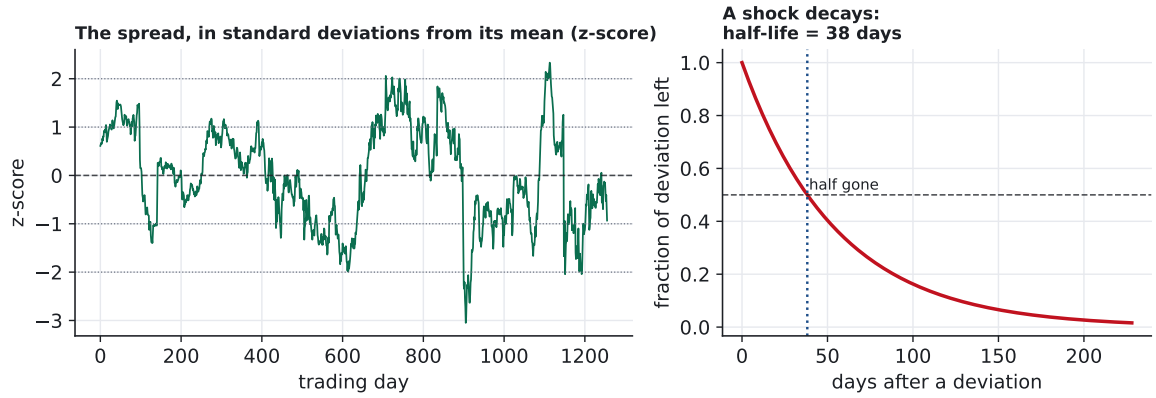

Figure 5: Turning the spread into a tradeable signal. **Left:** the Goldman/Morgan spread as a z-score—standard deviations from its mean—which is what the strategy actually watches. **Right:** the fitted mean-reversion speed implies a half-life of about 38 days; a typical dislocation loses half its size in that time and decays geometrically thereafter. The half-life sets the natural holding period of the trade.

Why a tethered spread must mean-revert: the error-correction view. There is a deeper reason a cointegrated spread reverts, and it has a name that will reappear in Part II: *error correction*. If two prices are cointegrated, then whenever the spread strays from its mean, at least one of the two prices must, on average, move to pull it back—otherwise the spread could not be stationary. We can write this dynamics out. Collecting the two prices, their changes obey

$$\begin{aligned} \Delta p_{1t} &= \gamma_1\,(s_{t-1} - \mu) + (\text{short-run terms}) + u_{1t}, \\ \Delta p_{2t} &= \gamma_2\,(s_{t-1} - \mu) + (\text{short-run terms}) + u_{2t}, \end{aligned} \tag{9}$$

where $s_{t-1} - \mu$ is last period's dislocation—the "error"—and the coefficients γ_1, γ_2 are the *speeds of adjustment* that "correct" it. For the spread to be stationary, the γ's must have signs that push it back: a positive dislocation (asset 1 rich) must drag p_1 down or p_2 up, or both. Equation (9) is the *error-correction model* (ECM), and it is the bridge from a static long-run

relationship (the spread has a mean) to the day-to-day dynamics a trader exploits (deviations get corrected). It also makes precise that cointegration and error correction are two sides of one coin—Granger's representation theorem, which we meet in its full multivariate form in Section 10 and Appendix C.

7 Putting on the trade

With a z-score signal and a sense of the reversion speed, the trading rule writes itself. Pick an entry threshold (say $|z| \geq 2$) and an exit threshold (say $z = 0$, the mean). When the z-score rises to $+2$ the spread is rich: *short* it—sell asset 1, buy β units of asset 2. When it falls to -2 the spread is cheap: *long* it. Close the position when the spread reverts to its mean. Figure 6 runs exactly this rule on Goldman/Morgan: the top panel marks every entry and exit against the bands, and the bottom panel accumulates the gross profit and loss.

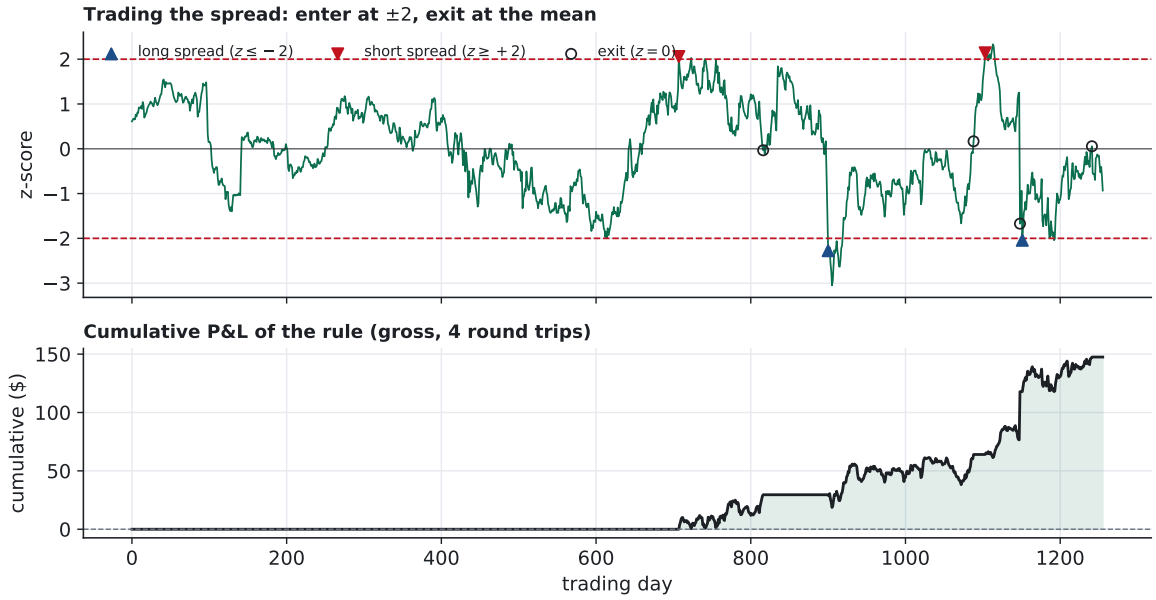

Figure 6: The rule in action on Goldman/Morgan. **Top:** the z-score with ±2 entry bands; long entries (△), short entries (▽) and exits at the mean (○) are marked. **Bottom:** cumulative gross P&L of holding the spread according to the rule. Over five years the two-standard-deviation rule fires only a handful of times—reversion this strong is rare—which is itself the real lesson: the edge is real but thin, and Section 12 and Appendix D discuss what eats it.

What the figure does and does not show. The cumulative line drifts up, which is encouraging—the spread did revert when dislocated, and the trades captured it. But three caveats are worth stating immediately, because they separate a teaching example from a deployable strategy. First, this is *gross* of costs; every round trip pays the *bid–ask spread* on both legs—the small, unavoidable gap between the price at which you can buy and the lower price at which you can sell, pocketed by the market maker, and not to be confused with the traded spread s_t—and for a thin edge those costs are decisive (Appendix D). Second, a ±2 rule on a single pair fires rarely—here only a few times in five years—so a real book trades *many* pairs at once to assemble a respectable number of independent bets. Third, and most important, the entire construction rests on the spread continuing to revert in the future as it did in the past—which is exactly the assumption that fails when a pair stops being cointegrated, the subject of Section 12.

The pairs trade, end to end. (1) Confirm the two prices are each $I(1)$ (ADF on the levels: a unit root). (2) Engle–Granger: regress to get the hedge ratio β; ADF the residual to confirm the spread is $I(0)$ (cointegrated). (3) Standardise the spread to a z-score; estimate the half-life to size the trade. (4) Enter when $|z|$ is large, exit when the spread reverts—then watch relentlessly for the tether breaking.

Part II — A whole basket

A single pair shows every idea, but real desks work with many assets at once: a basket of bank stocks, a set of currencies, a group of national equity indices. Two questions arise that a pair cannot pose. First, with n assets there may be *several* independent tethers among them—several cointegrating combinations—and we want to find all of them, not just one. Second, the assets influence one another dynamically: a move in one may lead moves in the others. Part II builds the multivariate machinery that handles both. The destination is the Johansen procedure and the vector error-correction model (Section 10); the vehicle that gets us there is the vector autoregression.

8 Many series at once: the vector autoregression

In plain terms. A vector autoregression (VAR) is the natural generalisation of "explain a variable by its own recent past" to a whole group of variables explaining *each other's* recent past. Instead of one equation, you have one equation per series, and each series is regressed on the recent history of itself and all the others. It is the standard workhorse for describing how a handful of markets move together and feed back on one another, and—more to our point—it is the scaffold inside which the cointegration machinery is built.

Stacking the n series into a vector $Y_t = (y_{1t}, \dots, y_{nt})^\top$, a VAR of order p is

$$Y_t = C + A_1 Y_{t-1} + A_2 Y_{t-2} + \cdots + A_p Y_{t-p} + \varepsilon_t, \tag{10}$$

where each A_k is an $n \times n$ matrix of coefficients (entry (i, j) says how yesterday's y_j feeds into today's y_i) and ε_t is a vector of zero-mean shocks. Each row is an ordinary regression, and the whole system is estimated equation by equation with least squares. Two practical choices arise: how many lags p to include, and—once fitted—how to read off the relationships. The lag count is chosen by an *information criterion* that trades goodness of fit against the number of parameters, penalising complexity so the model does not chase noise; the lag with the best score wins. (For the worked data in Section 11 this selects a single lag.) Reading off the relationships is the subject of the next section.

9 Who leads whom: Granger causality and impulse responses

A fitted VAR contains, buried in its coefficient matrices, the answer to "which markets lead which?". Two complementary tools extract it.

Granger causality: does one series help predict another? A series x is said to *Granger-cause* a series y if the past of x improves the forecast of y beyond what y's own past already provides. This is tested by fitting y's equation twice—once with the lags of x included, once without—and checking, with a standard F-test, whether the extra terms reduce the prediction error significantly. The name is a deliberate understatement: Granger causality is about

predictive *lead–lag* structure, not true cause and effect. If New York's move today helps forecast Hong Kong's tomorrow, New York "Granger-causes" Hong Kong in this purely forecasting sense—which may simply reflect time-zone ordering of the same global news, not any mechanism running from one to the other.

Impulse responses: how does a shock ripple out? Granger causality says *whether* one series leads another but not the *shape* of the influence—its sign, size or persistence. The *impulse response function* supplies that: it hits one series with a one-off unit shock and traces, through the fitted VAR, how every series responds over the following periods. Figure 7 does this for the six-index system of Section 11, shocking the US index (DJIA) by one unit on day zero. The other markets respond—the Hong Kong index most of all—and, because the selected model has a short memory, the ripple has essentially died out within two days. That rapid decay is itself the signature of a stationary dynamics: shocks fade, consistent with the differenced series being $I(0)$.

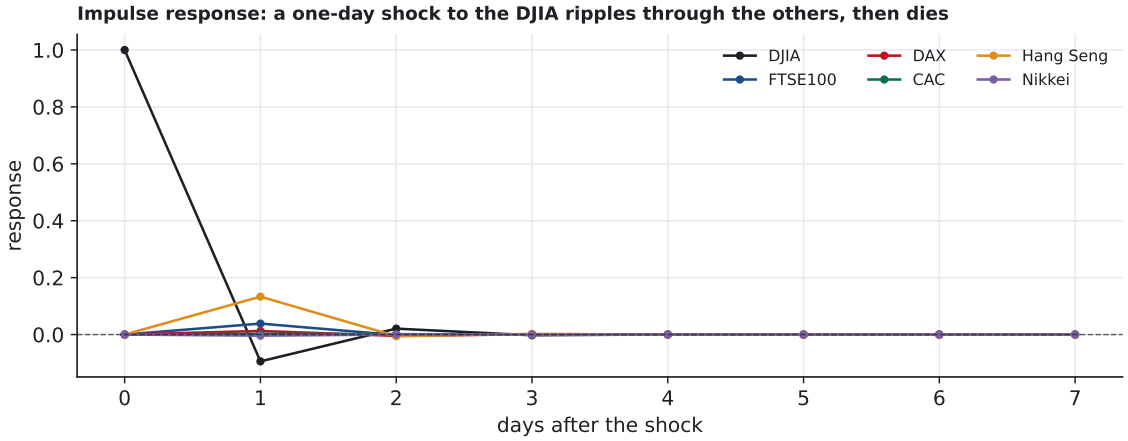

Figure 7: An impulse response from the fitted six-index VAR: a one-unit shock to the US index (DJIA) on day zero, and the response of all six markets over the following days. The reaction is led by the Hang Seng, and it dies away within about two days—shocks do not accumulate, which is what stationary (here, differenced) dynamics look like. Lead–lag tools like this map the short-run structure; cointegration (next section) captures the long-run tethers they miss.

These tools describe *short-run* dynamics—who reacts to whom over a few days. They are blind to the *long-run* tethers that pairs trading needs. A VAR fitted to differenced (stationary) data, in particular, throws the long-run relationships away entirely, because differencing erases the levels in which a tether lives. Recovering those tethers in the multivariate setting is the job of the Johansen procedure.

10 Johansen: finding every tether at once

In plain terms. With two assets there is at most one cointegrating combination, and Engle–Granger finds it. With n assets there can be up to $n-1$ independent tethers, and we want a method that finds *how many* there are and estimates *all of them* simultaneously, without having to pick one asset to put on the left of a regression. The Johansen procedure does exactly this. It is the multivariate, maximum-likelihood upgrade of Engle–Granger, and it is the standard tool for cointegration in a basket [2].

The starting point is to rewrite the VAR (10) in terms of the change ΔY_t and the lagged *level* Y_{t-1}—algebra that exactly generalises the rearrangement (3) we used for a single series.

The result is the **vector error-correction model** (VECM):

$$\Delta Y_t = \Pi\, Y_{t-1} + \sum_{k=1}^{p-1} \Gamma_k\, \Delta Y_{t-k} + \varepsilon_t. \tag{11}$$

Everything about the long run is concentrated in the single matrix Π that multiplies the lagged level. Its *rank* answers the cointegration question, by the following logic (Granger's representation theorem; Appendix C):

- If $\Pi = 0$ (rank 0), the levels never enter; the model is purely in differences and there are *no* tethers—no cointegration.
- If Π has *full* rank n, the levels themselves must be stationary, contradicting our assumption that each series is $I(1)$.
- In between, if Π has reduced rank r with $0 < r < n$, it factors as $\Pi = \alpha\beta^\top$ with β an $n \times r$ matrix whose r columns are the cointegrating vectors and α the matrix of speeds of adjustment. There are exactly r independent tethers, and $\beta^\top Y_t$ is a set of r stationary spreads.

So the entire problem reduces to determining the rank r of Π and recovering the factor β. Johansen's insight was that both fall out of an eigenvalue problem (the algebra is in Appendix C): each candidate tether comes with an eigenvalue measuring how strongly that combination reverts, and a sequence of tests on those eigenvalues—the *trace* and *maximum-eigenvalue* statistics—decides how many of them are real. The α in $\Pi = \alpha\beta^\top$ is the multivariate version of the speeds of adjustment γ_1, γ_2 from the pairwise error-correction model (9): it says, for each tether, which assets do the work of pulling the system back toward equilibrium.

Johansen, conceptually. Write the system as a VECM; all long-run information sits in the matrix Π on the lagged levels. The number of independent tethers is the rank of Π; the tethers themselves are read from its factorisation $\Pi = \alpha\beta^\top$ (columns of β = the stationary combinations, α = how fast each asset corrects). Rank 0: no cointegration. Reduced rank r: exactly r tradeable spreads, found all at once.

11 A worked example: six world equity indices

To see the whole Part II machine run on real data, we use a classic data set [10]: daily levels of six international equity indices—the US (DJIA), UK (FTSE 100), German (DAX), French (CAC 40), Hong Kong (Hang Seng) and Japanese (Nikkei) markets—rebased to 1.0 at the start and running over roughly two decades (about 5,300 trading days). Figure 8 shows them: six series that each wander like a random walk, with no shared visual level, yet which clearly travel in loose company.

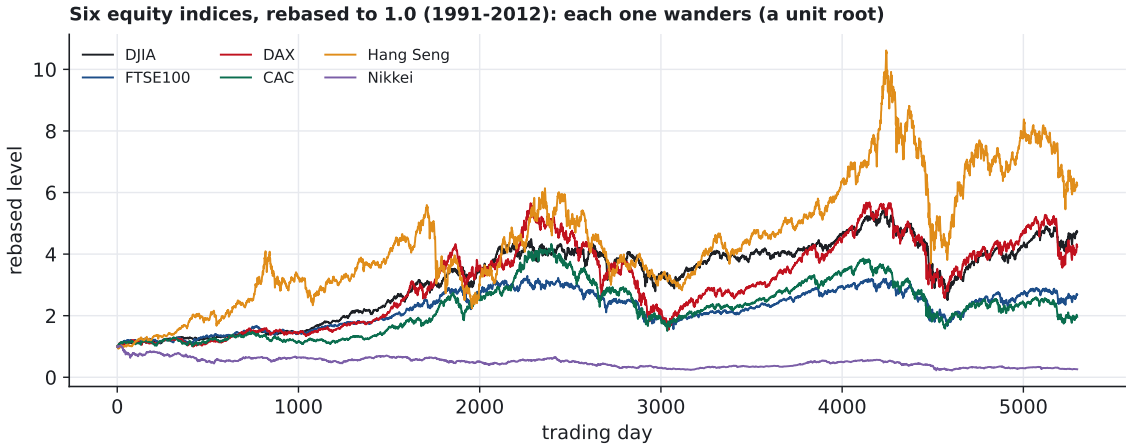

Figure 8: Six international equity indices, rebased to 1.0. Each one wanders (each is $I(1)$, as the ADF tests in Table 1 confirm), but they do not move independently—they are buffeted by common global forces. Johansen's question is how many stationary combinations of these six wandering series exist.

Step 1—are the individual series $I(1)$? We ADF-test each index in levels. Table 1 collects the statistics. Five of the six sit well above (less negative than) the 5% critical value of -3.41: we cannot reject a unit root, so they wander. Only the Nikkei, marginally, looks stationary in levels over this window. The five $I(1)$ series are exactly the raw material cointegration needs.

Table 1: Augmented Dickey–Fuller statistics for each index in levels (one lag, constant and trend). The 5% critical value is -3.41; a statistic above it fails to reject the unit root, i.e. the series wanders. Only the Nikkei is (marginally) stationary—the others are all $I(1)$.

	DJIA	FTSE100	DAX	CAC	Hang Seng	Nikkei
ADF statistic	-2.21	-2.31	-2.07	-1.43	-2.94	-3.95
unit root?	yes	yes	yes	yes	yes	no

Step 2—how many tethers? the Johansen rank test. We fit the VECM and run the trace test. Table 2 reports the eigenvalues (how strongly each candidate combination reverts) and the trace statistics against their 5% critical values. The test proceeds top down: the statistic for "rank ≤ 0" (187.1) far exceeds its critical value (94.2), rejecting no cointegration; "rank ≤ 1" and "rank ≤ 2" are likewise rejected; but at "rank ≤ 3" the statistic (21.5) finally falls below its critical value (29.7), and we stop. The data support a cointegrating rank of around $r = 2$–3—several independent stationary combinations among the six indices. Figure 9 visualises the leading one: the Johansen combination $\beta^\top Y_t$ is plainly stationary (it reverts around a fixed level), even though each ingredient index, like the single grey series shown for contrast, wanders.

Table 2: Johansen cointegration rank test on the six indices. Each row tests the null that the rank is at most r. The trace statistic exceeds its 5% critical value for $r = 0, 1, 2$ (reject) and falls below at $r = 3$ (stop)—evidence for two to three independent tethers. Eigenvalues measure the reversion strength of each successive combination.

null: rank $\leq r$	eigenvalue	trace statistic	5% critical	decision
0	0.01582	187.1	94.2	reject
1	0.00893	102.7	68.5	reject
2	0.00634	55.1	47.2	reject
3	0.00204	21.5	29.7	do not reject (stop)
4	0.00146	10.7	15.4	do not reject
5	0.00055	2.9	3.8	do not reject

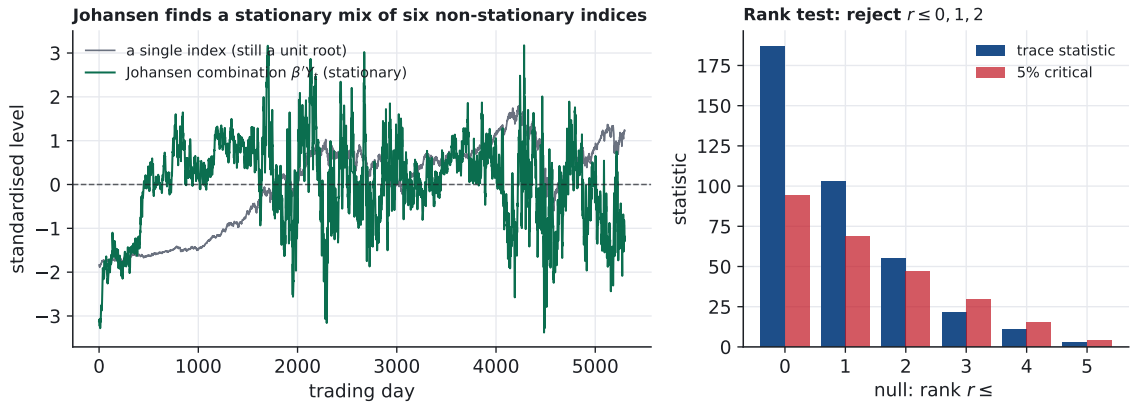

Figure 9: Johansen at work on the six indices. **Left:** the leading cointegrating combination $\beta^\top Y_t$ (green) is stationary—it reverts around a fixed level—whereas any single index (grey) wanders like a random walk; the procedure has manufactured a stationary series out of non-stationary parts. **Right:** the trace statistic exceeds its 5% critical value for ranks $0, 1, 2$, so the system has multiple tethers. Each green combination is, in principle, a tradeable spread.

What this buys a trader. Each of the r cointegrating combinations is a stationary spread—a basket-level analogue of the Goldman/Morgan spread—and each can be traded by exactly the Part I recipe: standardise to a z-score, enter on large dislocations, exit on reversion. Johansen has, in one pass, both certified that tradeable structure exists in the basket and handed over the specific weight vectors that isolate it. It is the multivariate engine behind the single-pair trade of Part I.

Part III — Where it breaks down

12 When the tether breaks

Everything so far rests on one assumption that deserves to be dragged into the light: that a relationship estimated on past data will persist into the future. Cointegration is a statement about the sample you measured. Markets are not obliged to honour it going forward, and when they stop, a pairs trade does not merely stop working—it can lose steadily and without any

signal to get out, because the strategy keeps insisting the spread will revert while it marches away.

The structural break. Figure 10 shows the failure mode that matters most. Two series are genuinely cointegrated for the first stretch of the sample—the spread, with its hedge ratio fitted on that in-sample window, sits quietly inside its band. Then something changes in the world (a merger, a regulatory shift, a change in business mix) and the relationship breaks. The hedge ratio that was correct before is wrong after; the spread leaves its band and never returns. A trader running the in-sample rule would have read the growing dislocation as an ever-better entry, added to the position, and been carried out. This is not a tail risk to be noted and dismissed—it is the central operational hazard of the entire approach.

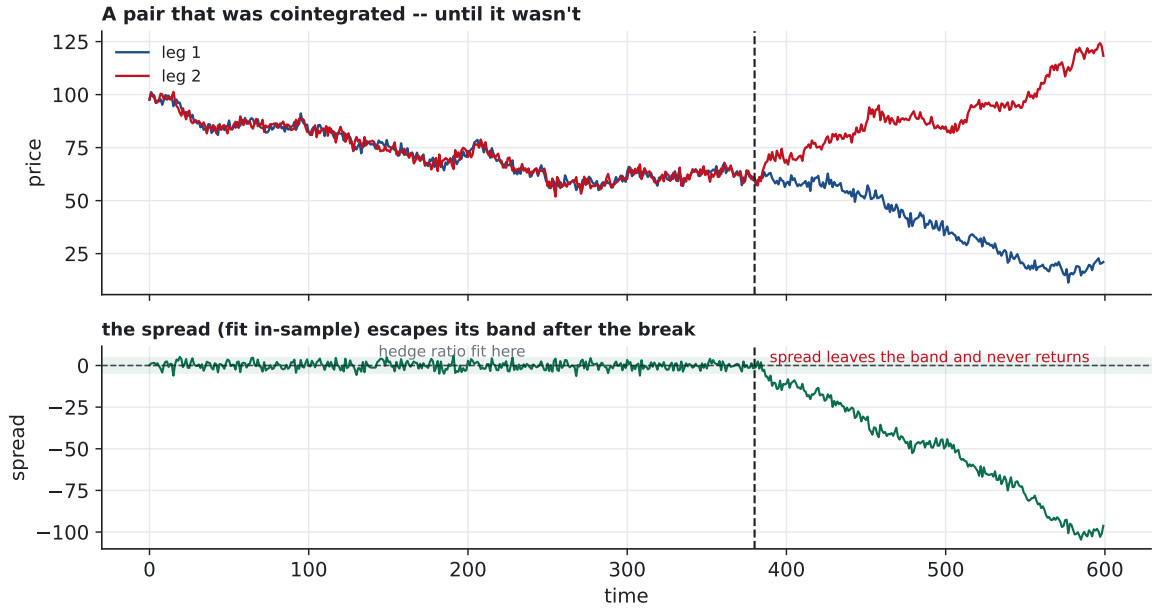

Figure 10: The real failure mode. A pair is cointegrated up to the dashed line (the spread, fit in-sample, stays in its band) and then suffers a structural break: the hedge ratio shifts and the spread escapes its band for good. A strategy that trusts the in-sample relationship reads the runaway dislocation as a stronger and stronger signal and loses without ever getting a reversion. No amount of in-sample testing prevents this—only out-of-sample discipline and a hard risk stop do.

The discipline this forces. Several habits follow directly, and they are what separate a backtest from a strategy.

- **Test out of sample.** A cointegration test on the same data you fitted is nearly circular. The real question is whether the spread *still* reverts on data the hedge ratio never saw—a walk-forward test, re-estimating β on a rolling window and checking reversion on the next.
- **Beware the spurious-regression trap.** Because a regression of one wandering series on another almost always looks significant (Appendix B), the ADF confirmation in step 2 of Engle–Granger is not optional dressing—it is the only thing standing between you and trading noise. Cointegration found without it is very often an artefact.
- **Respect the costs and the breadth.** A single pair throws off few signals and a thin per-trade edge (Section 7); after crossing the bid–ask on both legs the net can vanish (Appendix D). Real books diversify across many pairs to turn a small, repeatable edge into something worth trading—which only works if the edges are genuinely independent.

- **Put a hard stop on the spread.** Since a broken tether gives no reversion signal, the exit cannot rely on reversion alone. A risk limit—close if the z-score exceeds some larger level, or if the spread sits dislocated for longer than a few multiples of its half-life—is what caps the damage in Figure 10.

The one-sentence verdict. Cointegration tells you a tradeable spread existed *in the past*; only out-of-sample discipline, transaction-cost realism, and a hard risk stop can protect you when it stops existing in the *future*—and it eventually will.

13 Where this leaves us

The arc of this note is a single idea seen at two scales. A pairs trade wants a *stationary spread*—a combination of prices that reverts. A single price will not provide it, because a price is a random walk with a unit root and no level to revert to. *Cointegration* is the precise condition under which two (or many) such wandering series nonetheless admit a stationary combination, and the Augmented Dickey–Fuller test is how we tell a real tether from the spurious-regression mirage that levels data so readily produces. From the tether we built a z-score signal, a half-life to size the trade, and an error-correction view explaining why a tethered spread must revert at all. Scaling from a pair to a basket replaced the Engle–Granger regression with the Johansen procedure and the vector error-correction model, which find every independent tether at once—but changed none of the underlying logic. And running through everything is the discipline the cautionary section insists on: the relationships are properties of the past, the trade is taken in the future, and the gap between those two is where the risk lives. The time-series toolbox does not make that gap disappear; it makes it *visible*, measurable, and therefore manageable.

Part — Appendices

A The Dickey–Fuller distribution: why not the ordinary t-table

Section 4 warned that the unit-root test statistic does not follow the usual bell curve, so the familiar ± 1.96 cut-off is wrong. This appendix shows why—because the point is easy to state and important to internalise.

In an ordinary regression with well-behaved (stationary) data, a coefficient's t-ratio is approximately standard normal under the null that the coefficient is zero, which is where ± 1.96 comes from. The unit-root regression (3) violates the very assumption that result needs: under the null hypothesis of a unit root, the regressor y_{t-1} is itself a non-stationary random walk, not a stationary variable. The standard asymptotics break down, and the t-ratio converges instead to a non-standard distribution—expressible in terms of Brownian motion—first tabulated by Dickey and Fuller [3].

Figure 11 shows what this distribution actually looks like, obtained by simulating thousands of genuine random walks and computing the test statistic for each. The histogram is shifted markedly to the left of the standard normal and is not centred at zero but near -1.5. The practical consequence is stark: the true 5% critical value is about -2.86 (with a constant), far more negative than the normal's -1.64. If you used the normal cut-off you would reject the unit root—declare a series stationary, or a pair cointegrated—far more often than you should. In this simulation, where every series is a true random walk and the right answer is always "unit root", the normal cut-off would wrongly flag roughly 46% of them as stationary. Using the wrong table is, almost exactly, a machine for inventing tradeable relationships that do not exist.

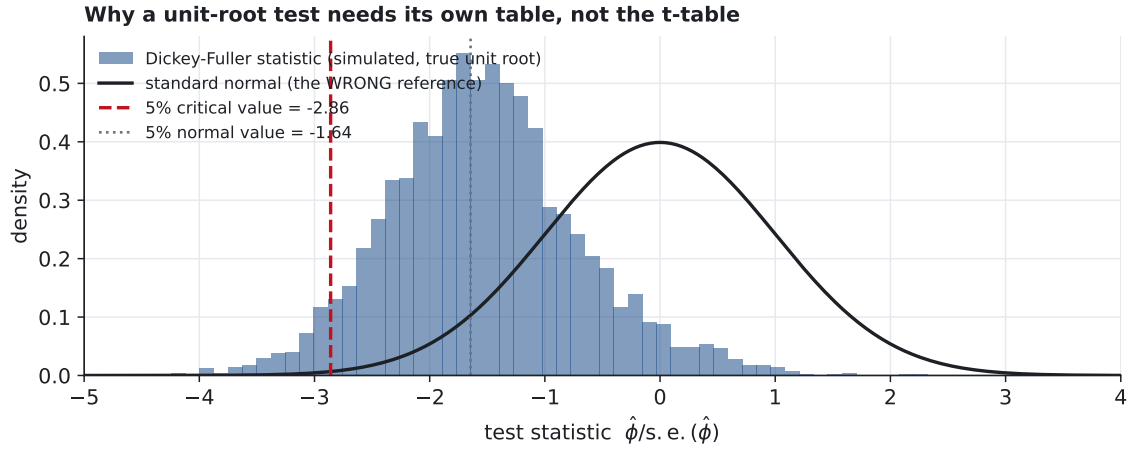

Figure 11: The Dickey–Fuller null distribution (blue histogram), simulated from thousands of true random walks, against the standard normal (black) one would wrongly use. The DF distribution is left-shifted and not centred at zero, so its 5% critical value (-2.86, red dashed) is far more negative than the normal's (-1.64, dotted). Judging a unit-root test with the normal table rejects the unit root far too readily—manufacturing spurious stationarity.

The same logic explains the footnote in Section 5.1: when the ADF test is applied to an *estimated* residual (the Engle–Granger spread), the distribution shifts further still, because the regression in step 1 actively chose β to make the residual look as stationary as possible. The Engle–Granger critical values (around -3.34 at 5% for two series) correct for that extra optimism, and grow more demanding as the number of series rises.

B Spurious regression: why levels regressions lie

Section 2 and the limitations section both lean on a striking fact: regressing one random walk on another *unrelated* random walk typically produces an impressive-looking, highly "significant" relationship that is entirely an illusion. This is the spurious-regression phenomenon, identified by Granger and Newbold [4] and explained theoretically by Phillips [5], and it is the precise danger that the ADF confirmation step in Engle–Granger exists to defuse.

Figure 12 demonstrates it by simulation. We generate thousands of pairs of *independent* random walks—by construction there is no relationship whatsoever—and regress one on the other. The left panel shows the distribution of the slope's t-statistic: instead of clustering inside ± 1.96 as it would for genuinely unrelated stationary variables, it spreads enormously, and about 86% of these meaningless regressions clear the ± 1.96 "significance" bar. The right panel shows the R^2: a median around 0.18 and frequently far higher, again from pure noise. The cause is the same non-stationarity as in Appendix A: two independent random walks will, by chance, both trend over any given sample, and a regression happily mistakes two unrelated trends for a relationship.

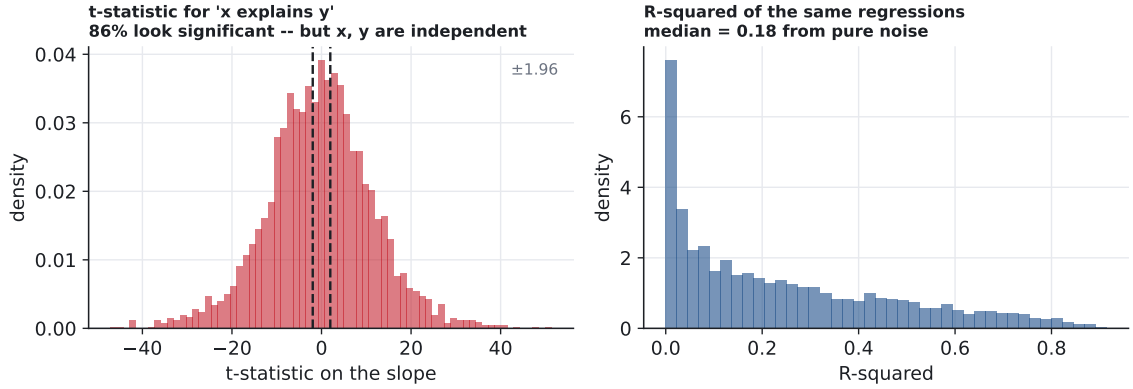

Figure 12: Spurious regression, simulated. Each of thousands of regressions is of one random walk on an *independent* other—no real relationship exists. **Left:** the slope t-statistic vastly overflows the ± 1.96 band ($\sim$86% "significant"). **Right:** R^2 values (median ~ 0.18) that look like real explanatory power but are pure noise. This is why a high-R^2, high-t regression of two prices means nothing on its own—and why cointegration must be confirmed by a unit-root test on the residual, not by the regression's apparent fit.

The moral is that the ordinary regression diagnostics—t-statistics, R^2—are simply not trustworthy when the variables are non-stationary levels. Cointegration is the disciplined exception: it is exactly the case where a levels regression is *not* spurious, and the only way to know you are in that exception is to test the residual for stationarity. Skip that test and you are, with overwhelming probability, trading the left panel of Figure 12.

C The vector error-correction model and Granger's theorem

This appendix fills in the algebra behind Section 10, for the reader who wants to see where the rank condition comes from.

From VAR to VECM. Start from the VAR (10) and subtract Y_{t-1} from both sides; with a little rearrangement (the multivariate version of turning $y_t = \phi y_{t-1} + \varepsilon_t$ into $\Delta y_t = (\phi - 1) y_{t-1} + \cdots$) one obtains the VECM (11), with the two kinds of coefficient matrices defined by

$$\Pi = \Big(\sum_{k=1}^{p} A_k \Big) - I, \qquad \Gamma_k = - \sum_{i=k+1}^{p} A_i. \tag{12}$$

The matrix Π multiplies the lagged *level* Y_{t-1} and so carries all the long-run information; the Γ_k multiply lagged *differences* and carry only short-run dynamics.

Granger's representation theorem. The key result is that, for a system of $I(1)$ variables, the rank of Π equals the number of cointegrating relationships, and a reduced rank r $(0 < r < n)$ is equivalent to the factorisation

$$\Pi = \alpha\,\beta^{\top}, \qquad \alpha, \beta \in \mathbb{R}^{n\times r}, \tag{13}$$

in which the columns of β are the cointegrating vectors (so $\beta^{\top}Y_t$ is r stationary spreads) and α holds the speeds of adjustment. The reason the factorisation must take this form is consistency of orders of integration: ΔY_t on the left of (11) is stationary ($I(0)$), and the short-run terms are stationary, so ΠY_{t-1} must be stationary too; but Y_{t-1} is $I(1)$, so Π can only act through directions that render it stationary—namely the r cointegrating combinations $\beta^{\top}Y_{t-1}$. Writing $\Pi = \alpha\beta^{\top}$ says precisely that Π first projects the levels onto those r stationary spreads ($\beta^{\top}Y_{t-1}$) and then feeds them back into each equation with weights α—the error-correction mechanism, now in matrix form. The scalar speeds γ_1, γ_2 of the pairwise model (9) are the entries of α in the two-series case.

Johansen's estimator. Johansen's procedure estimates this by maximum likelihood under the Gaussian assumption. After partialling the short-run terms ΔY_{t-k} out of both ΔY_t and Y_{t-1} by regression, leaving residual sets R_{0t} and R_{1t}, the cointegrating directions are the solutions of a generalised eigenvalue problem

$$\left|\, \lambda\, S_{11} - S_{10}\, S_{00}^{-1}\, S_{01} \,\right| = 0, \qquad S_{ij} = \frac{1}{T}\sum_{t} R_{it}R_{jt}^{\top}, \tag{14}$$

which yields eigenvalues $\lambda_1 > \cdots > \lambda_n$ (the reversion strengths in Table 2) and corresponding eigenvectors (the candidate β columns). The rank—how many eigenvalues are "really" non-zero—is decided by the trace and maximum-eigenvalue statistics

$$\begin{aligned} \lambda_{\text{trace}}(r) &= -T\sum_{i=r+1}^{n}\ln(1-\lambda_i), \\ \lambda_{\max}(r) &= -T\,\ln(1-\lambda_{r+1}), \end{aligned} \tag{15}$$

each compared against its own tabulated (non-standard, à la Appendix A) critical values. The trace statistic at rank r tests "at most r tethers" against "more than r"; one walks $r = 0, 1, 2, \ldots$ upward and stops at the first non-rejection, which is the procedure carried out in Table 2.

D Transaction costs and the thinness of the edge

Section 7 reported gross P&L and flagged that costs are decisive; this appendix says why in slightly more detail, without a full backtest.

Each round trip of a pairs trade crosses the bid–ask spread *four* times— buy and sell on each of the two legs—so the cost per round trip is roughly twice the combined half-spread of the pair. Against this, the gross edge per trade is the average distance the spread travels between entry and exit, which for an entry at $|z| = 2$ and exit at the mean is about two standard deviations of the spread. The trade is profitable only when that gross capture exceeds the round-trip cost, which gives a simple design rule: the entry threshold must be wide enough, relative to the bid–ask, that 2σ of spread comfortably clears four half-spreads of cost. Two forces then pull against each other. Widening the entry band raises the per-trade edge but fires less often, reducing the number of bets; narrowing it trades more but each trade clears less

above cost. Because a single pair already fires only a handful of times (Figure 6), the realistic path to a deployable strategy is *breadth*—many cointegrated pairs traded in parallel—so that a small, cost-surviving per-trade edge is repeated enough times to matter. This is also why the structural-break risk of Section 12 is so corrosive: it does not just zero out one pair's edge, it can turn the thin positive expectation negative, and across a diversified book the breaks need only be mildly correlated to undo the diversification the breadth was supposed to buy.

References

[1] R. F. Engle and C. W. J. Granger, *Co-integration and Error Correction: Representation, Estimation, and Testing*, Econometrica 55(2), 1987.

[2] S. Johansen, *Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models*, Econometrica 59(6), 1991.

[3] D. A. Dickey and W. A. Fuller, *Distribution of the Estimators for Autoregressive Time Series with a Unit Root*, Journal of the American Statistical Association 74(366), 1979.

[4] C. W. J. Granger and P. Newbold, *Spurious Regressions in Econometrics*, Journal of Econometrics 2(2), 1974.

[5] P. C. B. Phillips, *Understanding Spurious Regressions in Econometrics*, Journal of Econometrics 33(3), 1986.

[6] C. W. J. Granger, *Investigating Causal Relations by Econometric Models and Cross-spectral Methods*, Econometrica 37(3), 1969.

[7] J. D. Hamilton, *Time Series Analysis*, Princeton University Press, 1994.

[8] G. Vidyamurthy, *Pairs Trading: Quantitative Methods and Analysis*, Wiley, 2004.

[9] E. Gatev, W. N. Goetzmann and K. G. Rouwenhorst, *Pairs Trading: Performance of a Relative-Value Arbitrage Rule*, Review of Financial Studies 19(3), 2006.

[10] The six-index data set and the end-of-day cointegration pipeline are drawn from the author's CQF time-series project (`HFT_pairs`); the figures here recompute its ADF, Johansen and VECM results independently.