Pairs Trading with General State Space Models Structured and Explained Notes

Based on Zhang (2021)

Contents

1 Big Picture and Motivation

Pairs trading is a **market-neutral statistical arbitrage strategy** based on the assumption that two related assets share a long-run equilibrium relationship. When prices deviate temporarily, the trader sells the overpriced asset and buys the underpriced one, expecting mean reversion.

Traditional approaches suffer from two key limitations:

- They assume a **linear, Gaussian, homoscedastic** spread process.
- They rely on **static or ad-hoc trading rules** (e.g. $\pm 1\sigma$, $\pm 2\sigma$).

The contribution of the article is to jointly improve:

1. the **modeling of the spread**,
2. the **filtering of the latent spread**,
3. the **trading strategy**,
4. and the **choice of optimal trading rules**.

2 Modeling Framework: General State Space Model

2.1 Observation Equation

The observed prices are linked through a linear relationship:

$$P_{A,t} = \gamma P_{B,t} + x_t + \varepsilon_t, \tag{1}$$

where:

- γ is the hedge ratio,
- x_t is the **latent (true) spread**,
- $\varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ is observation noise.

2.2 State Equation: Spread Dynamics

The latent spread follows a general mean-reverting process:

$$x_{t+1} = f(x_t; \theta) + g(x_t; \theta), \eta_t, \tag{2}$$

This formulation allows for:

- **Nonlinearity** in mean reversion (f),
- **Heteroscedasticity** and volatility clustering (g),
- **Non-Gaussian innovations** (η_t).

2.3 Why This Matters

This general framework nests standard models (OU process) as special cases, but can also capture:

- fat tails,
- regime-dependent volatility,
- nonlinear pull-back forces.

This is crucial because trading performance depends on accurately modeling the *entire distribution* of the spread, not just its mean.

3 Notation Clarification: "Model (1)–(2)" vs "Model 1–5" vs "Model I–II"

3.1 What the paper calls Model (1) and Model (2)

The paper introduces a single general state-space framework consisting of:

$$P_{A,t} = \lambda + \gamma P_{B,t} + x_t + \varepsilon_t \tag{1}$$
$$x_{t+1} = f(x_t;\theta) + g(x_t;\theta)\,\eta_t \tag{2}$$

Equation **(1)** is the **observation equation** (how prices relate given a latent spread). Equation **(2)** is the **state equation** (the latent spread dynamics).

Identification issue (important). Because λ in (1) and the intercept term in $f(\cdot)$ (often called θ_1) cannot both be identified, the author sets $\lambda = 0$ and estimates the identifiable parameter vector

$$\psi = (\gamma, \theta, \sigma_\varepsilon).$$

So: (1)–(2) is the model. Everything else in the paper are specifications of (2).

3.2 Why do they later say "Model 1–Model 5"? (Section 3.5)

In **§3.5**, the author proposes a simulation-based method to choose optimal trading thresholds. To explain the method, he introduces **five example specifications of the state equation (2)** and labels them **Model 1, . . . , Model 5**. These are *not* new "overall" models; they are benchmark cases of (2) used for simulation experiments (comparing strategies A/B/C and showing how the Monte Carlo optimization works).

Concretely:

- Models 1–5 are illustrative data-generating processes for x_t (only equation (2) changes),
- equation (1) is conceptually still the observation equation in the framework,
- the point is: different choices of f, g, and η_t change the distribution and path properties of the spread, so the optimal (U, L) must be computed numerically.

3.3 Then why do they use "Model I" and "Model II" in Section 4?

In **§4 (Applications)**, the author tests the approach on real pairs and chooses two empirical specifications of the state equation (2), labelled **Model I** and **Model II**:

- **Model I (empirical)**: linear mean reversion + homoskedastic Gaussian innovation,

$$x_{t+1} = \theta_0 + \theta_1 x_t + \theta_2 \eta_t, \quad \eta_t \sim \mathcal{N}(0,1).$$

- **Model II (empirical)**: linear mean reversion + state-dependent volatility (heteroskedasticity),

$$x_{t+1} = \theta_0 + \theta_1 x_t + \big(\theta_2 + \theta_3 x_t^2\big)\eta_t, \quad \eta_t \sim \mathcal{N}(0,1).$$

So: Model I/II are the paper's chosen real-data cases of (2), while Models 1–5 are pedagogical/simulation cases of (2).

4 Filtering the Spread: Why and How

4.1 Why Filtering Is Needed

The spread x_t is not directly observable. Using raw residuals would mix:

- genuine mispricing signals,
- microstructure noise,
- observation error.

Filtering aims to extract the latent economic signal driving mean reversion.

4.2 Kalman Filtering Logic

At each time t, the filter:

1. predicts the next spread using the state equation,
2. updates this prediction using observed prices,
3. weights prediction vs. observation via the Kalman gain.

This produces a **smoothed, real-time estimate** $\bar{x}_t$ of the spread.

4.3 Why a Quasi Monte Carlo Kalman Filter (QMCKF)

Because the model is:

- nonlinear,
- non-Gaussian,

standard Kalman filters do not apply.
The article uses a **Quasi Monte Carlo Kalman Filter**:

- Gaussian mixture approximation for non-Gaussian shocks,
- low-discrepancy (Halton) sequences for numerical efficiency,
- recursive likelihood evaluation for MLE.

The output is:

- filtered spread $\bar{x}_t$,
- estimated parameters $\hat{\psi} = (\hat{\gamma}, \hat{\theta}, \hat{\sigma}_\varepsilon)$.

5 Trading Strategies

5.1 Benchmark Strategies (A and B)

All strategies rely on three thresholds:

- Upper boundary U,
- Lower boundary L,
- Mean level C.

Strategy A:

- Open when spread exceeds U or L,
- Close only when spread returns to the mean.

Strategy B:

- Always hold a position,
- Switch direction when boundaries are crossed.

5.2 New Strategy C (Key Innovation)

Strategy C modifies *when positions are opened and closed*:

- Open only when spread crosses boundaries *from outside*,
- Close when spread crosses the mean *or re-crosses boundaries*.

This avoids holding positions in extreme and volatile regimes.

5.3 Why Strategy C Works Better

- Lower drawdowns in homoscedastic settings,
- Higher returns *and* Sharpe ratios in heteroscedastic settings,
- Better alignment with volatility clustering.

6 Trading Rules: From Ad-Hoc to Optimal

6.1 Limitations of Classical Rules

Traditional rules ($\pm 1\sigma$, first-passage time, renewal theory):

- require linear Gaussian dynamics,
- break down under heteroscedasticity or nonlinearity.

6.2 Simulation-Based Optimal Rules (Section 3.5) — rigorous version

Why classical threshold rules fail here. Rules like $\pm 1\sigma$, first-passage-time, or renewal-theorem rules rely on strong structure (e.g. linear OU dynamics, homoskedastic Gaussian noise, tractable hitting-time distributions). Under the general spread dynamics (2) with heteroskedasticity, nonlinearity, or non-Gaussianity, those analytic tools break: even defining a constant "σ" boundary can be ill-posed when $g(x_t)$ varies over time.

Core idea. Given a chosen specification of (2) and estimated parameters $\hat{\psi}$ (from QMCKF+MLE), simulate many spread paths $\{x_t^{(n)}\}_{t=1}^{T}$ from (2), apply a candidate trading strategy (A/B/C) with candidate thresholds (U, L) (and typically $C = \mathbb{E}[x_t]$), compute performance, and choose the (U, L) that maximizes your objective.

Implementation details (exactly what the paper does). The paper normalizes thresholds in standard-deviation units:

$$U = \mu_x + u\,\sigma_x, \qquad L = \mu_x + \ell\,\sigma_x,$$

with grid

$$u \in \{0.1, 0.2, \ldots, 2.5\}, \qquad \ell \in \{-2.5, -2.4, \ldots, -0.1\}.$$

For each simulated path $n = 1, \ldots, N$ (paper uses $N = 10{,}000$) over $T = 1000$ trading days, and for each grid pair (u, ℓ), compute:

- cumulative return $CR_{u,\ell}^{(n)}$ under the chosen strategy,
- Sharpe ratio $SR_{u,\ell}^{(n)}$ under the chosen strategy,

then average across simulations:

$$CR_{u,\ell} = \frac{1}{N}\sum_{n=1}^{N} CR_{u,\ell}^{(n)}, \qquad SR_{u,\ell} = \frac{1}{N}\sum_{n=1}^{N} SR_{u,\ell}^{(n)}.$$

Finally select

$$(U^*, L^*) = \arg\max_{(u,\ell)} z_{u,\ell}, \qquad z \in \{CR,\, SR\}$$

(and similarly for other objectives like Calmar).

Transaction costs (important). The paper assumes 20bp per asset per transaction; since a pairs trade transacts *two* assets, a full round-trip trade implies 40bp total trading cost in the simulation experiments.

Why the paper introduces "Model 1–5" here. To *illustrate* the above procedure, the paper lists five benchmark specifications for the spread dynamics (2):

- Model 1: linear + Gaussian + homoskedastic (the classic case),
- Model 2: nonlinear mean reversion + Gaussian,
- Model 3: linear + Gaussian + heteroskedastic (state-dependent volatility),
- Model 4: linear + non-Gaussian (heavy-tailed t noise),
- Model 5: nonlinear + non-Gaussian.

These are not new overall frameworks: they are five different choices of (f, g, p) inside (2), used to compare how strategies A/B/C behave and how the optimal (U, L) changes.

Key lesson from the simulation table. The paper finds that when the spread is heteroskedastic (Model 3), Strategy C dominates the other strategies in *both* expected cumulative return and Sharpe ratio. For homoskedastic models, Strategy C often has competitive Sharpe but lower cumulative return (because it reduces time spent holding the position in extreme regions).

7 Empirical Applications and Results

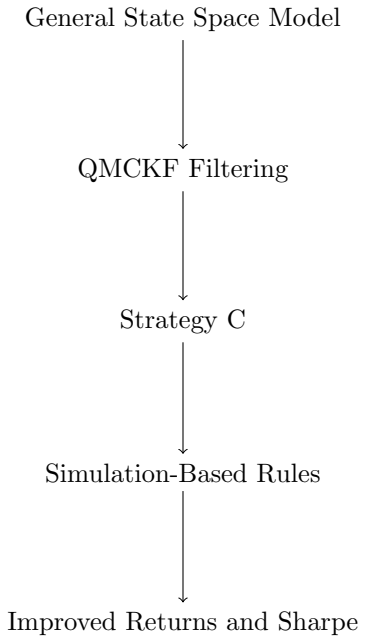
The methodology is applied to:

- Pepsi vs. Coca-Cola,
- Taiwan vs. Hong Kong ETFs,
- U.S. bank stocks (large vs. small).

Main findings:

- Strategy C + heteroscedastic model dominates benchmarks,
- Sharpe ratios improve dramatically,
- Smaller, more volatile assets yield better pairs trading performance.

8 Overall Logic of the Paper



9 Paper's Full Pipeline (Section 3.6, condensed)

1. **Choose a specification of (1)–(2)** (i.e., pick f, g, and the noise distribution for η_t).
2. **Estimate parameters and filter the spread** using QMCKF on $\{P_{A,t}, P_{B,t}\}_{t=1}^{T}$: obtain $\hat{\psi}$ and filtered spread $\{\bar{x}_t\}_{t=1}^{T}$.
3. **Choose a trading strategy** (A, B, or C), and **compute optimal thresholds** (U^*, L^*) by Monte Carlo simulation from the estimated spread dynamics (2) under $\hat{\psi}$.
4. **Out-of-sample trading:** for $t > T$, keep filtering $\bar{x}_t$ using $\hat{\psi}$ and apply the fixed strategy + (U^*, L^*) learned in-sample to generate trades.

10 Takeaway

Pairs trading performance depends jointly on modeling, filtering, strategy design, and rule optimization.

This article shows that once all four layers are treated consistently within a general state-space framework, statistical arbitrage performance improves substantially.