Pairs Trading a Sparse Synthetic Control

Jesus Villota

Introduction

Pairs Trading

Definition:

- Widely recognized as a cornerstone of ***statistical arbitrage***
- Exploits ***temporary divergences*** in the prices of two historically correlated or economically linked assets.

Modus Operandi

- ***Dollar-neutral*** strategy
- Implies simultaneously going:
  - **LONG** in the relatively **undervalued** asset
  - **SHORT** in the relatively **overvalued** one
- It aims to profit from the eventual ***convergence of prices*** or from ***mispricing correction***

Limitations

Pairs identification

- ***Difficult to identify*** pairs of economically linked assets
- ***High-dimensional asset pools*** exacerbate the problem

Time-varying correlations

- Even when a pair is identified, their ***relationship will change over time***

Sensitivity to noise

- Identifying mispricing in a pair is ***too sensitive to noise***
- Difficult to separate noise from mispricing

Non-linear dependencies

- Traditional approaches rely on ***distance*** measures or ***cointegration***-based criteria

Research Gap

Need for strategies that...

- Systematically identify latent economic linkages
- Mitigate overfitting in high-dimensional asset pools
- Robust to changing market conditions
- Successfully separate mispricing from noise
- Capture non-linear and tail dependencies

Research Question

Can the integration of ***sparse synthetic control*** with ***copula-based dependence modeling*** improve the performance of pairs trading strategies?

Our Approach

- Novel framework integrating:
  - **Sparse synthetic control methods**
  - **Copula-based dependence modeling**
- Designed to enhance:
  - ***Identification*** of pairs
  - Strategy ***adaptability***
  - ***Risk management*** capabilities
  - Performance ***stability***

Key Methodological Components

- **Sparse Synthetic Control**
  - Avoids reliance on a fixed or pre-specified partner asset
  - Constructs a *synthetic pair* as a sparse linear combination from a donor pool
- **Copula-Based Dependence Framework**
  - Captures non-linear relationships between the target & synthetic assets
  - Allows us to compute mispricing probabilities
- **Trading Signal Generation**
  - Based on relative mispricing between target and synthetic assets
  - Designed to filter market noise

Literature Review

Literature Review

Classical Pairs Trading

- **Gatev et al. (1998)**:
  - *First comprehensive academic study*
  - *Documented 11% annual excess returns (1962-2002)*
  - *Established distance-based methodology*
- **Elliott et al. (2005)**:
  - *introduced a mean-reverting Gaussian Markov chain model for spread dynamics*
  - *established analytical methods for parameter estimation using EM*

Empirical Validations

- **Chen et al. (2019)** *reports large abnormal returns driven by short-term reversals and pairs momentum effects*
- **Do et al. (2010)**: *showed that pairs trading remains viable in turbulent periods*
- **Krauss (2016) & Rad et al. (2016)**: *confirmed that distance, cointegration, and copula-based strategies can yield significant alpha but exhibit important differences regarding convergence speed and trading frequencies.*

Literature Review

Cointegration Approach

- **Vidyamurthy (2004)**: *seminal application to equity markets*
- **Caldeira and Moura (2013)**: *Brazilian market application*
- **Huck and Afawubo (2014)**: *cointegration outperforms distance methods*
- **Cartea and Jaimungal (2015)**: *Optimal dynamic investment strategies*
- **Lintilhac (2016)**: *Cryptocurrency applications*

Copula-Based Methods

- **Min and Czado (2010)**: *Bayesian inference for pair-copulas*
- **Stander et al. (2013)**: *Mispricing detection framework*
- **Liew and Wu (2013) & Xie et al. (2016)**: *Superior tail dependencies*
- **Krauss and Stubinger (2017) & Zhi et al. (2017)**: *Adaptive thresholds*
- **Recent extensions**:
  - Mixed copulas (da Silva et al., 2023)
  - ARMA-GARCH integration (Wang and Ding, 2023)
  - Cryptocurrency applications (Tadi and Witzany, 2025)

Literature Review

Advanced Modeling Techniques

- **Do (2006)**: *Stochastic residual spread models*
- **Zeng and Lee (2014)**: *Optimal threshold determination*
- **Johansson et al. (2024)**: *Convex-concave optimization for multi-asset statistical arbitrage*
- **Machine Learning Integration**:
  - *OPTICS clustering to constrain search space* (Sarmento and Horta, 2020)
  - *Reinforcement Learning for automated pairs selection* (Han et al., 2023)
  - *Graphical matching to reduce overlap among pairs* (Qureshi and Zaman, 2024)

Replication Methods

- **Alexander (1999), Alexander and Dimitriu (2002)**: *classical treatment connecting cointegration analysis and hedging tasks*
- **Alexander and Dimitriu (2005a, 2005b)**: *investigate how cointegration outperforms traditional techniques in crafting robust index trackers and exploiting time-varying market regimes*
- **Shu et al. (2020)**: *shows that sparse solutions across a large universe can reduce transaction costs*

- **Bradrania et al. (2021)**: *dynamic selection methods for index constituents using machine learning*

Research Positioning

Our contribution...

- Integrates sparse synthetic control with copula-based dependence modeling
- Overcomes the cumbersome pairs identification process
- Enhances adaptability to structural breaks and complex dependencies
- Provides systematic framework for high-dimensional asset pools

Methodology

Sparse Synthetic Control

- Let $\mathbf{y} = [y_t]_{t=1}^T$ be the log-price of target asset
- Let $\mathbf{X} = [x_{1t}, ..., x_{Nt}]_{t=1}^T$ be the log-prices of donor assets
- Build synthetic asset $\mathbf{y}^*$ as:

$$y_t^* = \sum_{i=1}^N w_i^* x_{it} \quad \text{for } t = 1, ..., T$$

- weights $\mathbf{w}^* = [w_1^*, ..., w_N^*]'$ solve:

$$\mathbf{w}^* = \arg\min_{\mathbf{w} \in \mathbb{R}^N} \left\{ \sum_{t=1}^T \left(y_t - \sum_{i=1}^N w_i x_{it} \right)^2 + \lambda \|\mathbf{w}\|_1 \right\} \quad \text{s.t.} \quad \mathbf{1}^\top \mathbf{w} = 1$$

Key Features

- ℓ_1 **Regularization (LASSO)**
  - Induces sparsity through non-differentiability at origin
  - Automatically selects most relevant assets
- **Optimization Properties**
  - Unique solution (convex problem)
  - Direct sparsity control via λ
  - Portfolio interpretation ($\mathbf{1}^\top \mathbf{w} = 1$)

- **Implementation**
  - Support identification via thresholding:

$$\mathcal{I} = \{i \in \{1, ..., N\} : |w_i^*| > \epsilon\}$$

  - Efficient solution via proximal algorithms

Empirical Application

- **Data**: S&P500 daily adjusted-close prices
- **Target**: NVIDIA (NVDA)
- **Period**: Jan 2010 - Jan 2025
- **Split**: 70% training, 30% testing
- **Result**: 27 stocks selected

Synthetic Control Model Weights

Ticker	Company	Weight (%)
AME	Ametek	41.08
LUV	Southwest Airlines	33.31
TFC	Truist Financial	25.60
AEP	American Electric Power	21.69
ADM		

Archer Daniels Midland

20.56

RSG

Republic Services

18.42

AXP

American Express

18.10

LLY

Lilly (Eli)

14.74

C

Citigroup

9.67

VRSN

Verisign

7.77

MTB

M&T Bank

7.38

FE

FirstEnergy

7.16

FIS

Fidelity National Info

5.21

PARA

Paramount Global

4.48

Ticker	Company	Weight (%)
TXT	Textron	2.21
STX	Seagate Technology	0.26
BIIB	Biogen	0.16
NFLX	Netflix	-1.04
FDX	FedEx	-2.39
UDR	UDR, Inc.	-3.95
V	Visa Inc.	-5.43
CNP	CenterPoint Energy	-7.75
MS	Morgan Stanley	

-16.21

NI

NiSource

-16.35

WMT

Walmart

-16.65

UNP

Union Pacific

-25.77

ADSK

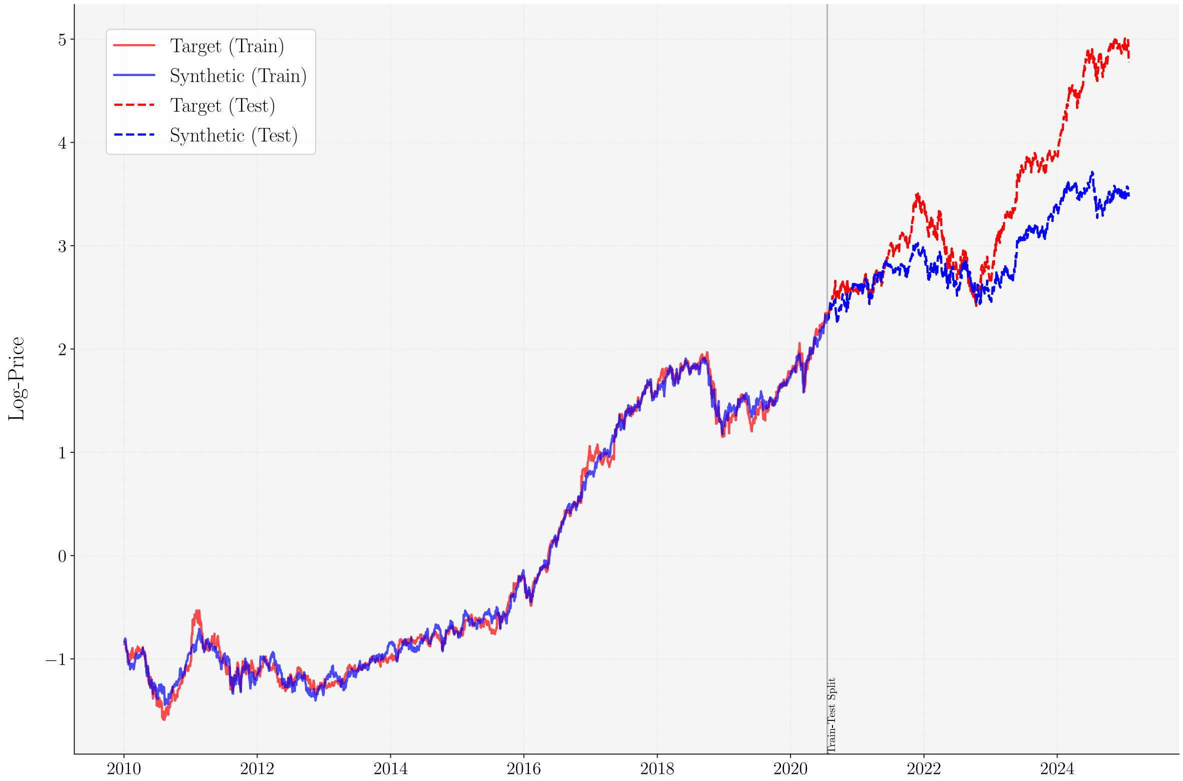
Autodesk

-42.25

Total

100.00

Target vs Synthetic Log-Prices



Copula-based Modeling

- Traditional pairs-trading approaches rely on *linear correlation* and *cointegration measures*
  - **Limitations**
    - Restrictive assumptions about joint distributions
    - Poor performance during market stress
    - Miss asymmetric tail dependencies
  - **Solution: Copula-based dependence modeling**
    - Decouples marginal distributions from joint dependence
    - Captures non-linear interactions
    - Allows to quantify mispricing probabilities

Sklar's Theorem

- Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.
- Let $R, R^* : \Omega \to \mathbb{R}$ be RVs representing ***target*** and ***synthetic*** log-returns.
- Let F_R and F_{R^*} denote their respective cumulative distribution functions (CDFs).
- Then, for the joint CDF F_{R,R^*}, there exists a copula $C : [0,1]^2 \to [0,1]$ s.t.:

$$F_{R,R^*}(r, r^*) = C(F_R(r), F_{R^*}(r^*)) \quad \forall r, r^* \in \mathbb{R}$$

- If F_R and F_{R^*} are continuous, then C is unique.
- Conversely, if C is a copula and F_R, F_{R^*} are CDFs, then F_{R,R^*} defined above is a joint CDF with margins F_R and F_{R^*}.

- When uniqueness holds, by the ***Probability Integral Transform***:

$$C(u, v) = \mathbb{P}(F_R(R) \leq u, F_{R^*}(R^*) \leq v) \quad \text{for} \quad (u, v) \in [0,1]^2.$$

- When it exists, the ***copula density*** $c : [0,1]^2 \to \mathbb{R}_+$ is given by

$$c(u, v) = \frac{\partial^2 C(u, v)}{\partial u \partial v},$$

- Then, the ***joint density*** can be decomposed as:

$$f_{R,R^*}(r, r^*) = c(F_R(r), F_{R^*}(r^*)) f_R(r) f_{R^*}(r^*)$$

Marginal Distribution Estimation

1. **Compute log-returns**:

$$r_t = y_t - y_{t-1} \quad \text{and} \quad r_t^* = y_t^* - y_{t-1}^*$$

2. **Estimate ECDFs**:

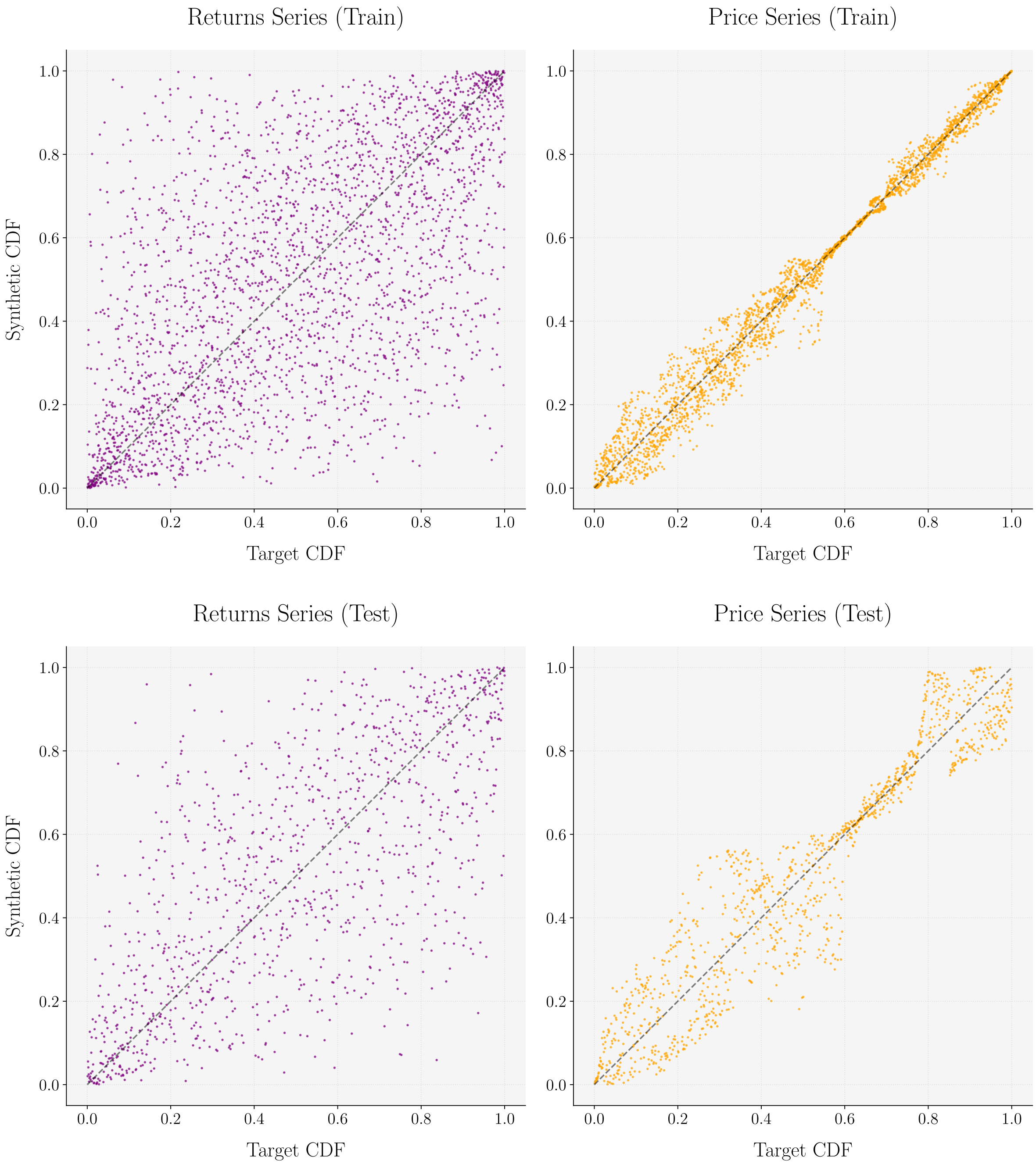
$$\hat{F}_R(r) = \frac{1}{T-1} \sum_{t=2}^{T} \mathbb{I}(r_t \leq r) \tag{1}$$

$$\hat{F}_{R^*}(r^*) = \frac{1}{T-1} \sum_{t=2}^{T} \mathbb{I}(r_t^* \leq r^*) \tag{2}$$

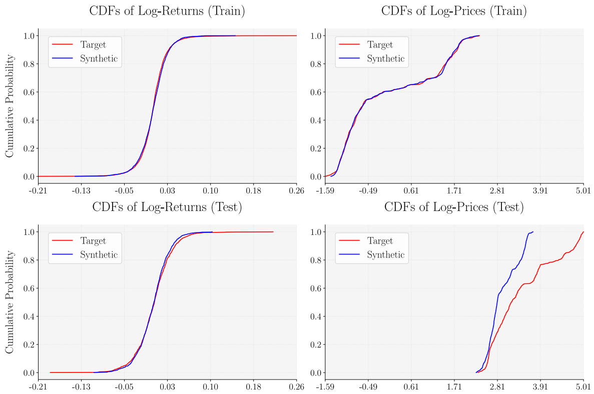
3. **Transform to uniforms**:

$$u_t = \hat{F}_R(r_t) \quad \text{and} \quad v_t = \hat{F}_{R^*}(r_t^*)$$

CDF Scatterplot: Returns Vs. Prices



Empirical CDFs: Returns Vs. Prices



Maximum Likelihood Estimation

For each copula family $\mathcal{C} = \{C_\theta : \theta \in \Theta\}$, estimate parameters via:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \ell(\theta|\mathbf{u}, \mathbf{v})$$

where the log-likelihood is:

$$\ell(\theta|\mathbf{u}, \mathbf{v}) := \sum_{t=2}^{T} \ln c_\theta(u_t, v_t)$$

and $c_\theta(u, v)$ is the copula density:

$$c_\theta(u, v) = \frac{\partial^2 C_\theta}{\partial u \partial v}(u, v)$$

Elliptical Copulas

Gaussian Copula: $\Theta = \{\rho \in (-1, 1)\}$

$$c_\rho^{Gauss}(u, v) = \frac{1}{\sqrt{1-\rho^2}} \exp\left(-\frac{\zeta_u^2 + \zeta_v^2 - 2\rho\zeta_u\zeta_v}{2(1-\rho^2)} + \frac{\zeta_u^2 + \zeta_v^2}{2}\right)$$

where $\zeta_u = \Phi^{-1}(u)$, $\zeta_v = \Phi^{-1}(v)$

Student-t Copula: $\Theta = \{\rho \in (-1, 1), \nu > 2\}$

$$c^t_{\rho,\nu}(u, v) = \frac{\Gamma(\frac{\nu+2}{2})\Gamma(\frac{\nu}{2})}{\sqrt{1-\rho^2}\Gamma(\frac{\nu+1}{2})^2} \frac{(1 + \frac{\zeta_u^2+\zeta_v^2-2\rho\zeta_u\zeta_v}{\nu(1-\rho^2)})^{-(\nu+2)/2}}{\prod_{i\in\{u,v\}}(1 + \frac{\zeta_i^2}{\nu})^{-(\nu+1)/2}}$$

where $\zeta_u = t_\nu^{-1}(u)$, $\zeta_v = t_\nu^{-1}(v)$

Archimedean Copulas

For generator function ψ_θ,

$$C_\theta(u, v) = \psi_\theta(\psi_\theta^{-1}(u) + \psi_\theta^{-1}(v))$$

Family

Parameter Range

Generator Function

Clayton

$\Theta = (0, \infty)$

$\psi_\theta(t) = (1 + t)^{-1/\theta}$

Gumbel

$\Theta = [1, \infty)$

$\psi_\theta(t) = \exp(-t^{1/\theta})$

Frank

$\Theta = \mathbb{R} \setminus \{0\}$

$\psi_\theta(t) = -\frac{1}{\theta} \ln(1 - (1 - e^{-\theta})e^{-t})$

Joe

$\Theta = [1, \infty)$

$\psi_\theta(t) = 1 - (1 - e^{-t})^{1/\theta}$

Mixed Copulas

- N14: Rotated Clayton-Gumbel mixture with $\Theta \subset \mathbb{R}^2_+$

Characterization of Copulas

Elliptical Copulas

- **Gaussian Copula**
  - Symmetric dependence
  - Light tails
- **Student-t Copula**
  - Symmetric dependence
  - Heavy tails

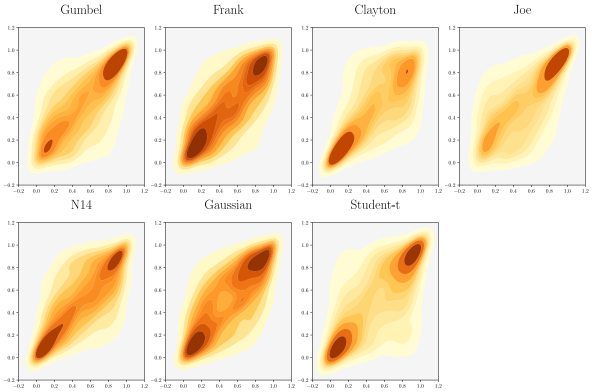
Archimedean Copulas

- **Clayton**
  - Asymmetric, lower tail dependence
- **Gumbel**
  - Asymmetric, upper tail dependence
- **Joe**
  - Asymmetric, strong upper tail dependence
- **Frank**
  - Symmetric, no tail dependence

Mixed Copula (N14)

- Asymmetric tail dependence

Calibrated Copula Density Heatmaps



Model Selection

Information Criteria:

$$\begin{aligned} \text{AIC} \quad &= 2k - 2\ell(\widehat{\theta}|\mathbf{u},\mathbf{v}) \\ \text{SIC} \quad &= k\ln(T-1) - 2\ell(\widehat{\theta}|\mathbf{u},\mathbf{v}) \\ \text{HQIC} \quad &= 2k\ln(\ln T - 1) - 2\ell(\widehat{\theta}|\mathbf{u},\mathbf{v}) \end{aligned}$$

Key Findings:

- Student-t copula provides best fit, followed by N14, Gaussian and Frank
- Elliptical copulas provide a better fit than Archimedean copulas
- Heavy tails are significant
- Symmetric dependence structure dominates
- Tail dependence is more relevant than symmetry (N14 > Gaussian, Frank)

Copula Fitting Results

Copula

SIC

AIC

HQIC

Joe

-671.50

-677.39

-675.26

Clayton

-1168.92

-1174.80

-1172.67

Gumbel

-1210.02

-1215.90

-1213.78

Frank

-1212.68

-1218.56

-1216.43

Gaussian

-1337.69

-1343.57

-1341.44

N14

-1425.18

-1431.06

-1428.94

Student-t

-1427.05

-1432.94

-1430.81

Trading Strategy

Pairs Trading via Mispricing Indices

- We adapt the mispricing index (MI) strategy from Xie et al. (2016) to our setting
- We trade a target asset (with returns R_t) against its synthetic counterpart (with returns R_t^*).
- While the strategy might initially appear unconventional, it hinges on interpreting conditional probabilities of daily returns as an evolving measure of relative mispricing.

Mispricing Index (MI)

- Two conditional mispricing indices
- $MI_t^{R|R^*}$ ⇒ *How "mispriced" is the* ***target asset*** *today conditional on today's* ***synthetic return****?*

$$MI_t^{R|R^*} := \mathbb{P}(R_t \leq r_t \mid R_t^* = r_t^*) = \frac{\partial C_{\hat{\theta}}(F_R(r_t), F_{R^*}(r_t^*))}{\partial F_{R^*}(r_t^*)}$$

- $MI_t^{R^*|R}$ ⇒ *How "mispriced" is the* ***synthetic asset*** *today conditional on today's* ***target return****?*

$$MI_t^{R^*|R} := \mathbb{P}(R_t^* \leq r_t^* \mid R_t = r_t) = \frac{\partial C_{\hat{\theta}}(F_R(r_t), F_{R^*}(r_t^*))}{\partial F_R(r_t)}$$

Cumulative Mispricing Index (CMI)

- Individual MI reflects only instantaneous view
- Solution:
  - Accumulate signals over time to track persistent mispricing
  - Reset them to zero after position is closed to prevent stale signals
- This defines a **Cumulative Mispricing Index** (CMI) for each asset:

$$\mathrm{CMI}_t^R = \begin{cases} \mathrm{CMI}_{t-1}^R + (MI_t^{R|R^*} - 0.5), & \text{if no position reset at time } t, \\ 0, & \text{if a position is closed at } t, \end{cases} \tag{3}$$

$$\mathrm{CMI}_t^{R^*} = \begin{cases} \mathrm{CMI}_{t-1}^{R^*} + (MI_t^{R^*|R} - 0.5), & \text{if no position reset at time } t, \\ 0, & \text{if a position is closed at } t. \end{cases} \tag{4}$$

where $CMI_0^R = CMI_0^{R^*} = 0$.

Trading Logic - Signal Generation

- **"Or-Or" Logic**: Proposed by Xie et al. (2016)
  - Trades are initiated when either asset is shows mispricing
  - Positions are closed when either asset corrects
- **"And-Or" Logic**: Proposed by Rad et al. (2016)

  - Requires concurrent signals from both assets to open positions
  - Mispricing correction in either asset triggers position closure
  - This logic is more conservative & yields *more robust performance*

- **Parameters**:
  - Entry thresholds: $(D_l, D_u) = (-0.6, 0.6)$
  - Stop-loss boundaries: $(S_l, S_u) = (-2, 2)$

Trading Rule

Trading Rule given the current CMIs (CMI_t^R, $\mathrm{CMI}_t^{R^*}$) and previous signal (TR_{t-1}):

$$TR_t(\mathrm{CMI}_t^R, \mathrm{CMI}_t^{R^*}, TR_{t-1}; D_l, D_u, S_l, S_u) =$$

$$\begin{cases} +1 & \text{if } (\mathrm{CMI}_t^R \le D_l \text{ and } \mathrm{CMI}_t^{R^*} \ge D_u) \\ -1 & \text{if } (\mathrm{CMI}_t^R \ge D_u \text{ and } \mathrm{CMI}_t^{R^*} \le D_l) \\ 0 & \text{if } \left\{ \begin{array}{l} \left\{ TR_{t-1} = 1 \quad \text{and } [(\underbrace{\mathrm{CMI}_t^R \ge 0 \text{ or } \mathrm{CMI}_t^{R^*} \le 0}_{\text{take profit}}) \text{ or } (\underbrace{\mathrm{CMI}_t^R \le S_l \text{ or } \mathrm{CMI}_t^{R^*} \ge S_u}_{\text{stop loss}})] \right\}, \text{or} \\ \left\{ TR_{t-1} = -1 \text{ and } [(\underbrace{\mathrm{CMI}_t^R \le 0 \text{ or } \mathrm{CMI}_t^{R^*} \ge 0}_{\text{take profit}}) \text{ or } (\underbrace{\mathrm{CMI}_t^R \ge S_u \text{ or } \mathrm{CMI}_t^{R^*} \le S_l}_{\text{stop loss}})] \right\} \end{array} \right. \\ TR_{t-1} & \text{otherwise} \end{cases}$$

Position Entry and Exit Conditions

- **Long target/Short synthetic (+1)**:
  - Target *underpriced* ($\mathrm{CMI}_t^R \le D_l$) **AND** Synthetic *overpriced* ($\mathrm{CMI}_t^{R^*} \ge D_u$)

- **Short target/Long synthetic (-1)**:
  - Target *overpriced* ($\mathrm{CMI}_t^R \ge D_u$) **AND** Synthetic *underpriced* ($\mathrm{CMI}_t^{R^*} \le D_l$)

- **Exit position (0)**: Triggered by either:
  - **Take profit**: Either CMI crosses zero (*price correction*)
  - **Stop loss**: Either CMI exceeds stop-loss boundaries

Strategy Implementation

1. **Daily process**:
   - Obtain returns for target (r_t) and compute synthetic returns (r_t^*)
   - Transform to uniform margins: $u_t = \hat{F}_R(r_t)$, $v_t = \hat{F}_{R^*}(r_t^*)$
   - Compute MIs using fitted copula partial derivatives
   - Update CMIs based on previous values and exit conditions
   - Generate trading signal based on CMI thresholds

2. **Position management**:
   - Dollar-neutral portfolio (equal capital in long and short)
   - Reset CMIs after closing positions
   - Exit positions based on either take-profit or stop-loss

Operational Requirements

1. Access to **basket trading** capabilities

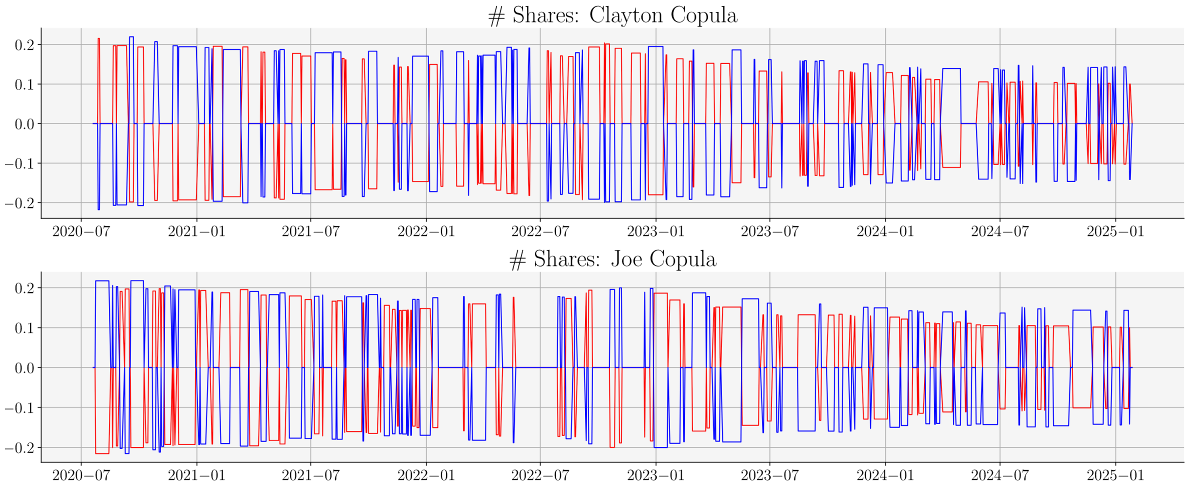
- *Available to institutional investors*
- *Modern execution systems treat 27 components as single basket order*
- *Reduced transaction costs through optimized order routing*

2. Sufficient **liquidity** in all components

- *donor pool is restricted to S&P 500 universe*

Results

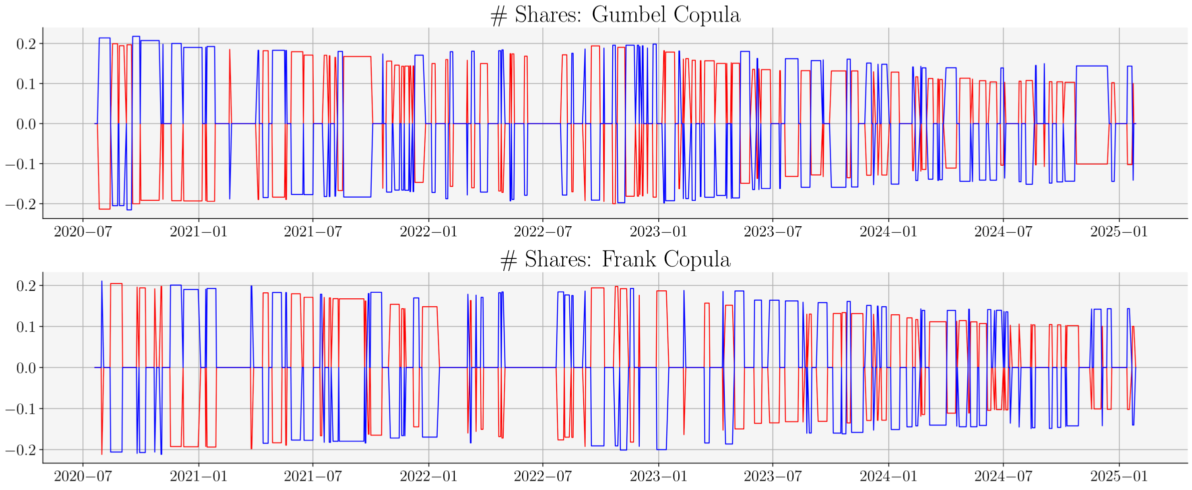
Position Sizes Over Time



Shares: Clayton Copula

Shares: Joe Copula

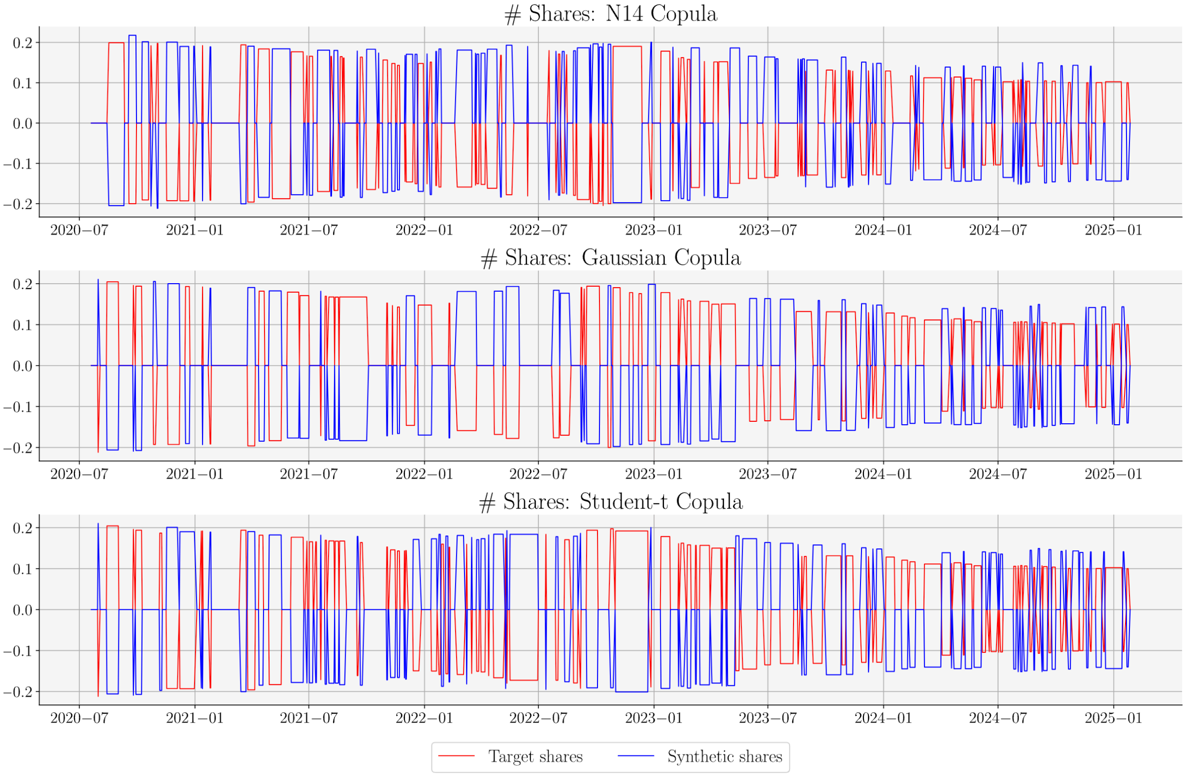
Position Sizes Over Time



Shares: Gumbel Copula

Shares: Frank Copula

Position Sizes Over Time



Performance Hierarchy

- **Top tier**: N14 mixed copula (~78% return)
- **Middle tier**: Student-t, Clayton, Gumbel (~73-75% return)
- **Lower tier**: Joe, Frank (~67% return)
- **Lagging**: Gaussian (~63% return)

Performance Metrics by Copula

Copula

Total Return (%)

Ann. Return (%)

Ann. Vol. (%)

Sharpe Ratio

Sortino Ratio

Calmar Ratio

Max DD (%)

VaR-95 (%)

N14

77.82

17.26

4.35

3.97

5.75

11.25

1.53

-0.32

Clayton

74.67

16.56

4.18

3.97

5.30

10.89

1.52

-0.31

Student-t

74.63

16.55

4.60

3.60

4.64

7.70

2.15

-0.33

Gumbel

72.59

16.10

4.61

3.49

4.42

7.42

2.17

-0.35

Joe

67.45

14.96

4.62

3.24

3.85

5.83

2.57

-0.36

Frank

66.53

14.76

3.97

3.71

4.75

10.85

1.36

-0.30

Gaussian

62.70

13.91

4.43

3.14

4.10

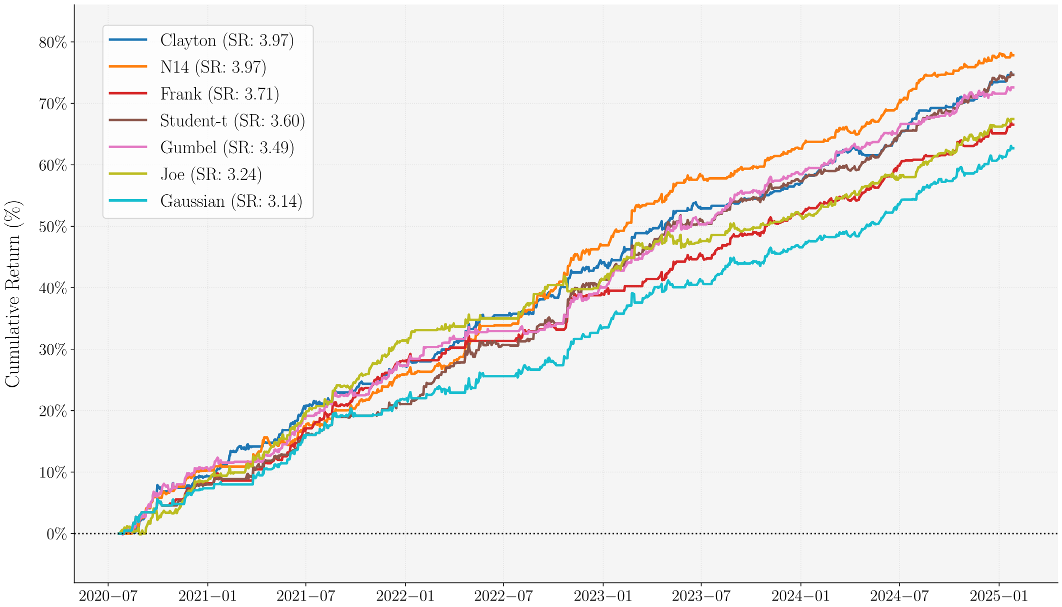
8.12

1.71

-0.35

*All metrics computed over out-of-sample period from July 2020 to January 2025

Equity Curves Comparison


Clayton (SR: 3.97)
N14 (SR: 3.97)
Frank (SR: 3.71)
Student-t (SR: 3.60)
Gumbel (SR: 3.49)
Joe (SR: 3.24)
Gaussian (SR: 3.14)
Cumulative Return (%)