Pairs-Trading a Sparse Synthetic Control

Jesus Villota Miranda[†][*]

⟨ [†]CEMFI, Calle Casado del Alisal, 5, 28014 Madrid, Spain ⟩
⟨ Email: jesus.villota@cemfi.edu.es ⟩

This version: 13th February 2025

Abstract

Financial markets frequently exhibit transient price divergences between economically linked assets, yet traditional pairs trading strategies struggle to adapt to structural breaks and complex dependencies, limiting their robustness in dynamic regimes. This paper addresses these challenges by developing a novel framework that integrates sparse synthetic control with copula-based dependence modeling to enhance adaptability and risk management. Economically, our approach responds to the need for strategies that systematically identify latent linkages while mitigating overfitting in high-dimensional asset pools. The sparse synthetic control methodology constructs a parsimonious synthetic asset via a constrained linear combination of candidates from a broad donor pool, automating pair selection while prioritizing interpretability and computational efficiency. By embedding this within a copula-based dependence framework, we capture non-linear and tail dependencies between target and synthetic assets. Trading signals, grounded in the relative mispricing between these assets, employ a cumulative index that resets after position closures to isolate episodic opportunities, with disciplined entry rules requiring concurrent misalignment signals to filter noise. Empirical analysis demonstrates the superior performance of our approach across diverse market conditions.

JEL Codes: C14, C32, C58, C61, G12, G14

Keywords: Pairs Trading, Sparse, Synthetic Control, High leverage, Dollar neutral, Copula,

[*] I gratefully acknowledge financial support from the Santander Research Chair at CEMFI.

1. Introduction

Pairs trading is widely recognized as a cornerstone of statistical arbitrage, offering a market-neutral investment approach that exploits temporary divergences in the prices of historically correlated or economically linked assets. By simultaneously taking a long position in the relatively undervalued asset and a short position in the relatively overvalued one, pairs traders aim to profit from the eventual convergence of these prices. This strategy has garnered enduring prominence among quantitative researchers and practitioners, attributing its appeal to both conceptual simplicity–focusing on the relative mispricing of two assets–and the potential for stable returns independent of broader market movements.

While pairs trading is conceptually straightforward, its effective implementation faces notable complexities in practice. Traditional approaches often rely on simple distance measures or cointegration-based criteria to identify pairs and establish entry and exit rules. However, these methods can be hampered by strict parametric assumptions, sensitivity to transient noise, and an inability to adapt to evolving market conditions. Structural breaks, non-linear dependencies, and time-varying correlation patterns often violate the assumptions of classical linear models, increasing the risk of identifying spurious relationships and making it difficult to achieve stable performance over diverse market regimes.

To address these challenges, recent research has explored more flexible frameworks that combine advanced econometric tools with statistical learning. In particular, incorporating synthetic control methodologies and copula-based dependence modeling aims to better capture the dynamic interactions between assets. By abandoning the sole reliance on fixed, potentially fragile pair relationships, such approaches promise to more robustly uncover the underlying economic or statistical linkages that drive temporary mispricings, thus laying the groundwork for improved performance and risk control in pairs trading strategies.

Building on the challenges and limitations outlined above, this paper proposes a novel pairs trading framework that integrates sparse synthetic control methods with copula-based dependence modeling. The primary research question we aim to answer is: "*Can the integration of sparse synthetic control and copula-based dependence modeling improve the performance of pairs trading strategies?*" To address this question, we design a methodology that overcomes several shortcomings of traditional pairs trading.

First, rather than relying on a fixed or pre-specified partner asset, we construct a *synthetic asset* through a sparse linear combination of assets from a larger donor pool. This allows the framework to discover the most influential contributors to the target asset's behavior, effectively automating pair selection. By enforcing sparsity in the weight vector, we reduce computational complexity and enhance interpretability, while mitigating overfitting risks in thinner markets.

Second, we incorporate copula-based dependence modeling to capture potentially complex, non-linear relationships and tail dependencies that can arise in financial returns. Unlike correlation- or cointegration-based strategies, which often impose strict distributional assumptions, copulas decouple the marginal distributions from the joint dependence structure, thereby offering a more nuanced view of how assets co-move. This feature is especially important in periods of market stress, when returns frequently exhibit heightened correlations and non-linearities.

Finally, we adapt and extend the Mispricing Index (MI) strategy of Xie et al. (2016) by introducing a Cumulative Mispricing Index (CMI) that resets upon trade closure, ensuring that stale signals do not accumulate across different trading episodes. As in Rad et al. (2016), we adopt an "*and-or*" logic for opening and closing positions, requiring persistent mispricing signals from both the target and synthetic assets to initiate a trade and closing positions promptly when either market correction or stop-loss conditions are met.

The remainder of this paper proceeds as follows. In Section 1 we begin by reviewing the relevant literature on pairs trading, synthetic control methods, and copula-based dependence modeling. In Section 2 we present our methodological framework, detailing how sparse synthetic control and copula families are jointly employed to construct a robust trading signal, and introduce the mispricing index (MI) strategy adapted to incorporate copula-driven signals. Subsequently, in Section 3 we conduct an empirical evaluation using real-world market data, illustrating the performance and practical implications of our approach. We conclude in Section 4 by summarizing key insights, discussing limitations, and outlining prospective directions for future research.

1.1 Literature Review

Pairs trading has emerged as a cornerstone of statistical arbitrage strategies, with a rich history in both academic research and practical applications. The foundational work of Gatev et al. (2006) provided the first comprehensive academic study of pairs trading, documenting significant excess returns of up to 11% annually for self-financing portfolios over a 40-year period from 1962 to 2002. This seminal paper was complemented by the theoretical framework developed in Elliott et al. (2005), which introduced a mean-reverting Gaussian Markov chain model for spread dynamics and established analytical methods for parameter estimation using the EM algorithm.

Empirical investigations have thoroughly examined the profitability of pairs trading across different markets and time periods. For instance, Chen et al. (2019) reported large abnormal returns driven by short-term reversals and pairs momentum effects, while Do and Faff (2010) showed that simple pairs trading remains viable in turbulent periods despite a general profitability decline in later years. In a UK-centric study, Bowen and Hutchinson (2014) recorded moderate annual returns once risk and liquidity were accounted for. Large-scale assessments in Krauss (2016) and 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.

A popular way to identify and exploit persistent relationships in pairs trading has involved cointegration analysis. Vidyamurthy (2004) stands out as a seminal reference, detailing how cointegration can be applied to detect mean-reverting spreads in equity markets. Subsequent research has explored various aspects of this approach: Caldeira and Moura (2013) demonstrated the effectiveness of cointegration-based selection methods in the Brazilian market, while Huck and Afawubo (2014) provided evidence that cointegration-based strategies outperform distance-based methods. Cartea and Jaimungal (2015) extended the framework by incorporating optimal dynamic investment strategies, and Lintilhac and Tourin (2016) applied these techniques to cryptocurrency markets.

A growing strand of research leverages copulas to model more general dependencies beyond linear correlation. Min and Czado (2010) introduced Bayesian inference for multivariate copulas using pair-copula constructions, while Stander et al. (2013) offer a copula-based approach for detecting relative mispricing. Extensions in Liew and Wu (2013) and Xie et al. (2016) underscore that copulas outperform distance-of-prices rules in capturing tail dependencies Multi-dimensional variants have been proposed (e.g., Lau et al. (2016)) to incorporate three or more assets into a single framework. Further refinements, like those introduced in Krauss and Stübinger (2017) and Zhi et al. (2017), combine t-copulas or dynamic copula-GARCH models with individualized thresholds for improved risk-adjusted returns. In the high-frequency domain, Chu and Chan (2018) showed that copula-based mispricing indices can be coupled with deep learning for profitability enhancements. Recent efforts also explore mixed copulas (Sabino da Silva et al. (2023)), ARMA-GARCH approaches (Wang and Ding (2023)), and copulas specialized for cointegrated assets (He et al. (2024)), culminating in improved alpha extraction. Finally, Tadi and Witzany (2025) proposes reference-asset-based copula trading specifically for cryptocurrencies.

Practical guidance and pedagogical discussions on pairs trading can be found in Joubert et al. (2021), which provides a broad compendium of methods, from classical cointegration to machine learning-based selection. On a methodological note, Alexander (2008) offers valuable introductions to both cointegration analysis and copula applications in financial markets, particularly in chapters II.5 and II.6.

Beyond cointegration or copula methodologies, several innovative techniques have surfaced. Do et al. (2006) developed a stochastic residual spread model, while Zeng and Lee (2014) focused on optimal threshold determination. In more recent research, Sarmento and Horta (2020) incorporates machine learning (OPTICS clustering) to constrain search space, while Johansson et al. (2024) leverages convex-concave optimization for multi-asset statistical arbitrage. Reinforcement learning is featured in Han et al. (2023) for automated pair selection, and Qureshi and Zaman (2024) employs a graphical matching approach to reduce overlap among chosen pairs. Further, Roychoudhury et al. (2023) couples clustering with deep RL for equity indices, whereas Rotondi and Russo (2025) applies a partial correlation-based distance to cluster promising trading candidates.

The method of replicating a target asset's returns by constructing a portfolio of contributor assets is reminiscent of index-tracking procedures. Classic treatments connecting cointegration analysis and hedging tasks (e.g., Alexander (1999) and Alexander and Dimitriu (2002)) lay theoretical groundwork for such an approach. Subsequent refinements in Alexander and Dimitriu (2005a) and Alexander and Dimitriu (2005b) investigate how cointegration outperforms traditional techniques in crafting robust index trackers and exploiting time-varying market regimes. Complementary research (e.g., Shu et al. (2020)) shows that sparse solutions across a large universe can reduce transaction costs, an idea further corroborated in Bradrania et al. (2021), where machine learning identifies dynamic selection methods for index constituents. These frameworks illustrate how synthetic control concepts provide a flexible foundation for building market-neutral positions or tracking assets with fewer assumptions.

2. Methodology

2.1 Sparse Synthetic Control

The core component of our pairs trading strategy involves constructing a synthetic asset that replicates the price behavior of a target security (e.g: AAPL) using a combination of assets from a donor pool. Let $\mathbf{y} = [y_t]_{t=1}^T \in \mathbb{R}^T$ denote the log-price time series of a target asset and $\mathbf{X} = [x_{1t}, ..., x_{Nt}]_{t=1}^T \in \mathbb{R}^{T \times N}$ denote the log-price time series of a donor pool of assets. We construct a synthetic asset $\mathbf{y}^*$ through a sparse linear combination

$$y_t^* = \sum_{i=1}^N w_i^* x_{it}.$$

The weights $\mathbf{w}^* = [w_1^*, ..., w_N^*]$ are determined via a cardinality-constrained quadratic program

$$\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathbb{R}^N} \sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{N} w_i x_{it} \right)^2 \quad \text{s.t.} \quad \left| \begin{array}{ll} \mathbf{1}^\top \mathbf{w} & = 1 \\ \|\mathbf{w}\|_0 & \leq K \end{array} \right.$$

where $\|\mathbf{w}\|_0 := \sum_{i=1}^{N} \mathbb{I}(w_i \neq 0)$ counts the non-zero elements in $\mathbf{w}$. The goal is to enforce sparsity so that only a limited number of assets receive a nonzero weight. The NP-hard cardinality constraint is approximated by the following procedure:

1. Solve the full least squares problem

$$\mathbf{w}^{(1)} = \arg\min_{\mathbf{w}\in\mathbb{R}^N} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 \quad \text{s.t.} \quad \mathbf{1}^\top \mathbf{w} = 1.$$

2. Select the K largest weights (in absolute value) from $\mathbf{w}^{(1)}$ into

$$\mathcal{I} := \{ i : |w_i^{(1)}| \text{ among } K \text{ largests} \}$$

3. Solve the restricted program on support $\mathcal{I}$

$$\mathbf{w}^{(2)} = \arg\min_{\mathbf{w}_\mathcal{I}\in\mathbb{R}^K} \|\mathbf{y} - \mathbf{X}_\mathcal{I}\mathbf{w}_\mathcal{I}\|_2^2 \quad \text{s.t.} \quad \mathbf{1}^\top \mathbf{w}_\mathcal{I} = 1$$

   where $\mathbf{X}_\mathcal{I} \in \mathbb{R}^{T\times K}$ is the resricted donor matrix and $\mathbf{w}_\mathcal{I} \in \mathbb{R}^K$ is the restricted weight vector for the selected assets.

4. Construct the full weight vector $\mathbf{w}^* \in \mathbb{R}^N$ by embedding the optimized restricted weights back into the original N-dimensional space.

$$w_i^* = \begin{cases} w_j^{(2)} & \text{if } i = \mathcal{I}_j \\ 0 & \text{otherwise} \end{cases}$$

3. Copula-Based Dependence Modeling

The sparse synthetic control framework provides an adaptive mechanism to construct a replicating portfolio that dynamically identifies influential assets from a broad candidate pool. However, the efficacy of a pairs trading strategy depends not only on accurate synthetic replication but also on quantifying how –and to what extent– the target and synthetic assets co-move under varying market conditions. Traditional pairs-trading approaches often rely on linear correlation or cointegration measures, but these methods impose restrictive assumptions about the joint distribution of returns. Such assumptions are frequently violated in practice, particularly during periods of market stress where asymmetric tail dependencies and non-linear dynamics dominate.

To overcome these limitations, we complement the synthetic asset construction with copula-based dependence modeling. Copulas provide a flexible framework to decouple marginal distributions from the joint dependence structure, enabling us to capture non-linear and tail-dependent interactions that linear correlations overlook, model time-varying dependencies without assuming Gaussianity or stationarity and quantify conditional mispricing probabilities in a distributionally robust manner. We now formalize the copula framework and its integration with the synthetic asset returns.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $R, R^* : \Omega \to \mathbb{R}$ be real-valued random variables representing the target and synthetic log-returns, respectively, Let F_R and F_{R^*} denote their respective cumulative distribution functions (CDFs).

Definition 1 (Copula)**.** *A bivariate copula is a function $C : [0,1]^2 \to [0,1]$ satisfying:*

1. $C(u,0) = C(0,v) = 0$ *and* $C(u,1) = u$, $C(1,v) = v$ *for all* $u, v \in [0,1]$ *(boundary conditions)*
2. $C(u_2,v_2) - C(u_2,v_1) - C(u_1,v_2) + C(u_1,v_1) \geq 0$ *for all* $u_1 \leq u_2$, $v_1 \leq v_2$ *in* $[0,1]$ *(2-increasing)*

The fundamental relationship between copulas and joint distributions is established by Sklar's theorem:

Theorem 1 (Sklar (1959))**.** *Let F_{R,R^*} be the joint CDF of (R, R^*). Then there exists a copula $C : [0,1]^2 \to [0,1]$ such that*

$$F_{R,R^*}(r, r^*) = C(F_R(r), F_{R^*}(r^*)) \quad \forall r, r^* \in \mathbb{R}. \tag{1}$$

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, the copula can be expressed through the probability integral transform:

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

The corresponding copula density $c : [0,1]^2 \to \mathbb{R}_+$, when it exists, is given by $c(u,v) = \frac{\partial^2 C(u,v)}{\partial u \partial v}$, and the joint density can be expressed as $f_{R,R^*}(r,r^*) = c(F_R(r), F_{R^*}(r^*)) f_R(r) f_{R^*}(r^*)$, where f_{R,R^*} is the joint density and f_R and f_{R^*} are the marginal densities.

Intuitively, Sklar's theorem tells us that any joint distribution can be decomposed into two parts: the marginal distributions of individual variables and a copula that captures their dependence structure. This decomposition provides a framework for modeling the dependence structure between the target and synthetic returns independently of their marginal distributions. The implementation involves three stages: (1) nonparametric estimation of the marginal CDFs F_R, F_{R^*} , (2) copula calibration from parametric classes $\mathcal{C} = \{C_\theta : \theta \in \Theta\}$ via maximum likelihood estimation, (3) selection of an appropriate copula family

3.1 Marginal Distribution Estimation

The foundation of copula modeling lies in the accurate estimation of marginal distributions for both target and synthetic asset returns. To maintain flexibility and avoid restrictive parametric assumptions, we adopt a non-parametric approach through empirical cumulative distribution functions (ECDFs).

First, we construct logarithmic return series for both assets. Let y_t and y_t^* denote the log-prices of the target and synthetic assets at time t, respectively. The log-returns are computed as $r_t = y_t - y_{t-1}$ and $r_t^* = y_t^* - y_{t-1}^*$ for $t = 2, \ldots, T$, delivering return time series $\{r_t\}_{t=2}^T$ and $\{r_t^*\}_{t=2}^T$ for the target and stationary assets respectively.

Next, we estimate the marginal distributions through linearly interpolated ECDFs. For any $r \in \mathbb{R}$, the empirical distribution functions are given by

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

where $\mathbb{I}(\cdot)$ denotes the usual indicator function. Following Joubert et al. (2021), we then enforce linear interpolation between observed returns to ensure continuity of the distribution functions across their support. Also, to mitigate numerical instabilities during subsequent copula estimation, we constrain the ECDF outputs within $[\epsilon, 1-\epsilon]$ where $\epsilon = 10^{-5}$, thereby avoiding boundary effects at the distribution tails.

The final step involves applying the probability integral transform to obtain uniform marginals. Specifically, we compute pseudo-observations

$$u_t = \hat{F}_R(r_t) \quad \text{and} \quad v_t = \hat{F}_{R^*}(r_t^*) \quad \text{for } t = 2, \ldots, T,$$

yielding paired realizations $(\mathbf{u}, \mathbf{v}) = \{(u_t, v_t)\}_{t=2}^T$ that reside in the unit square $[0,1]^2$. This transformation, justified by Sklar's Theorem, effectively decouples the marginal distributions from the dependence structure. The resulting uniform variates serve as canonical inputs for copula specification while preserving the essential dependence characteristics between target and synthetic returns.

3.2 Copula calibration from parametric classes

The goal of copula fitting is to find the copula that best describes the dependence structure between the returns of the target and synthetic assets. This is done by maximizing the likelihood of the observed data under different copula models. We consider parametric copula families $\mathcal{C} = \{C_\theta : \theta \in \Theta\}$ where each copula C_θ has density $c_\theta(u,v) = \frac{\partial^2 C_\theta}{\partial u \partial v}(u,v)$. For each candidate copula family, we estimate parameters via constrained maximum likelihood:

$$\hat{\theta} = \arg\max_{\theta\in\Theta} \ell(\theta|\mathbf{u},\mathbf{v}) \quad \text{where} \quad \ell(\theta|\mathbf{u},\mathbf{v}) := \sum_{t=2}^{T} \ln c_\theta(u_t, v_t). \tag{2}$$

The optimization is subject to parameter constraints Θ specific to each copula family:

- **Elliptical Copulas:**
  - Gaussian: $\Theta = \{\rho \in (-1,1)\}$ with density

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

    where $\zeta_u = \Phi^{-1}(u)$, $\zeta_v = \Phi^{-1}(v)$ and Φ is the standard normal CDF.
  - Student-t: $\Theta = \{\rho \in (-1,1), \nu > 2\}$ with density

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

    where $\zeta_u = t_\nu^{-1}(u)$, $\zeta_v = t_\nu^{-1}(v)$ and t_ν is the Student-t CDF.
- **Archimedean Copulas:** For generator function ψ_θ,

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

  - Clayton: $\Theta = (0,\infty)$ with $\psi_\theta(t) = (1+t)^{-1/\theta}$
  - Gumbel: $\Theta = [1,\infty)$ with $\psi_\theta(t) = \exp(-t^{1/\theta})$
  - Frank: $\Theta = \mathbb{R} \setminus \{0\}$ with $\psi_\theta(t) = -\frac{1}{\theta}\ln\left(1-(1-e^{-\theta})e^{-t}\right)$
  - Joe: $\Theta = [1,\infty)$ with $\psi_\theta(t) = 1-(1-e^{-t})^{1/\theta}$
- **Mixed Copulas:**
  - N14: Rotated Clayton-Gumbel mixture with $\Theta \subset \mathbb{R}_+^2$

A formal description of the copula fitting procedure can be found in Algorithm 2.

3.3 Selection of an appropriate copula family

After estimating parameters for each candidate copula family $\mathcal{C} = \{C_\theta : \theta \in \Theta\}$, we select the optimal model using information criteria that balance goodness-of-fit against model complexity. Let $\ell(\hat{\theta}|\mathbf{u}, \mathbf{v}) = \max_{\theta\in\Theta} \sum_{t=2}^{T} \ln c_\theta(u_t, v_t)$ be the maximized log-likelihood for a copula with parameter estimate $\hat{\theta}$, where T is the sample size and k is the number of parameters. We evaluate the following information criterions:

$$\begin{array}{lll} \textit{Akaike} & \text{AIC} & = 2k - 2\ell(\hat{\theta}|\mathbf{u}, \mathbf{v}) \\ \textit{Schwarz/Bayesian} & \text{SIC} & = k\ln(T-1) - 2\ell(\hat{\theta}|\mathbf{u}, \mathbf{v}) \\ \textit{Hannan-Quinn} & \text{HQIC} & = 2k\ln(\ln T - 1) - 2\ell(\hat{\theta}|\mathbf{u}, \mathbf{v}) \end{array}$$

The copula family with the lowest value for a chosen criterion is selected as optimal. These criteria penalize overfitting through the k term while rewarding better fit through the log-likelihood.

Table 1: Copula Model Selection Criteria

Copula	SIC	AIC	HQIC
Student-*t*	-138.78	-145.51	-143.18
Gumbel	-73.85	-80.58	-78.25
Clayton	-53.64	-60.37	-58.04
Joe	-50.61	-57.34	-55.01
Gaussian	-44.21	-50.93	-48.60
Frank	-38.37	-45.10	-42.76
N14	-	-	-

Table 1 presents the fitting results for different copula families.

4. Pairs Trading Strategy via Mispricing Indices (MI)

In this section, we adapt the mispricing index (MI) strategy from Xie et al. (2016) to our setting, wherein 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. Below, we detail the essential components of the approach and how trading positions are opened and closed.

4.1 Mispricing Index (MI), Flags and Cumulative Mispricing Index (CMI)

On each trading day t, let r_t and r_t^* respectively denote the realized returns for the target and synthetic assets. We define two conditional mispricing indices,

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

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

The quantity $MI_t^{R|R^*}$ measures how "mispriced" the target asset appears when conditioned on that day's synthetic return, whereas $MI_t^{R^*|R}$ does the same for the synthetic asset when conditioned on the target return. Since a single day's mispricing index reflects only an instantaneous view, we accumulate daily signals over time to gauge how much the returns have gradually driven prices apart (or together). We define a *flag* series for each asset, defined as a running sum of daily deviations from 0.5[1]. Let $\text{Flag}_R(0) = \text{Flag}_{R^*}(0) = 0$, then, for $t = 1, ..., T$ we have

$$\begin{aligned} \text{Flag}_t^R &= \text{Flag}_{t-1}^R + (MI_t^{R|R^*} - 0.5) &= \textstyle\sum_{s=1}^t (MI_s^{R|R^*} - 0.5), \\ \text{Flag}_t^{R^*} &= \text{Flag}_{t-1}^{R^*} + (MI_t^{R^*|R} - 0.5) &= \textstyle\sum_{s=1}^t (MI_s^{R^*|R} - 0.5). \end{aligned}$$

Similar to plotting cumulative returns, these raw flags track the net effect of mispricing signals over time.

To prevent the compounding of stale mispricing signals, we formally define a Cumulative Mispricing Index (CMI) as the reset-adjusted flag series through the recursive relationship:

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

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

where $\text{CMI}_0^R = \text{CMI}_0^{R^*} = 0$. Unlike the raw flags that accrue continuously, each CMI absorbs daily mispricing signals only until a trade is exited, at which point it is reset to zero. This mechanism ensures that any fresh mispricing accumulates from a "clean slate," thereby preventing the influence of past, already-traded mispricing from compounding future signals.

We formally present the procedures to compute the mispricing index and update the cumulative mispricing indices in Algorithm 3. and Algorithm 4.

[1] The subtraction of 0.5 centers the cumulative sum so that deviations from zero reflect mispricing.

4.2 Trading Logic

We implement a dollar-neutral trading strategy that capitalizes on relative mispricing signals between the target and synthetic assets. The trading rule (TR) we employ builds upon the frameworks of Xie et al. (2016) and Rad et al. (2016), incorporating their key insights about signal combination logic. While Xie et al. (2016) originally proposed an "*or-or*" framework, where trades are initiated when either asset shows mispricing and closed when either asset exhibits correction, Rad et al. (2016) demonstrated that a more conservative "*and-or*" approach yields more robust performance. This latter approach requires concurrent mispricing signals from both assets to open positions while maintaining a sensitive exit strategy where correction in either asset triggers position closure.

Let D_l and D_u denote the lower and upper thresholds for opening positions, and S_l and S_u the lower and upper stop-loss boundaries. Starting with $TR_0 = 0$, for $t = 1, ..., T$, the trading rule evolves as follows:

$$TR_t(\text{CMI}_t^R, \text{CMI}_t^{R^*}, TR_{t-1}; D_l, D_u, S_l, S_u) = \tag{3}$$
$$\begin{cases} +1 & \text{if } (\text{CMI}_t^R \leq D_l \text{ and } \text{CMI}_t^{R^*} \geq D_u) \\ -1 & \text{if } (\text{CMI}_t^R \geq D_u \text{ and } \text{CMI}_t^{R^*} \leq D_l) \\ 0 & \text{if } \left\{ \begin{array}{l} \left\{ TR_{t-1} = 1 \quad \text{and } \left[\underbrace{(\text{CMI}_t^R \geq 0 \text{ or } \text{CMI}_t^{R^*} \leq 0)}_{\text{take profit}} \text{ or } \underbrace{(\text{CMI}_t^R \leq S_l \text{ or } \text{CMI}_t^{R^*} \geq S_u)}_{\text{stop loss}} \right] \right\}, \text{or} \\ \left\{ TR_{t-1} = -1 \text{ and } \left[\underbrace{(\text{CMI}_t^R \leq 0 \text{ or } \text{CMI}_t^{R^*} \geq 0)}_{\text{take profit}} \text{ or } \underbrace{(\text{CMI}_t^R \geq S_u \text{ or } \text{CMI}_t^{R^*} \leq S_l)}_{\text{stop loss}} \right] \right\} \end{array} \right. \\ TR_{t-1} & \text{otherwise} \end{cases}$$

That is, at the beginning of each trading day t, observe the current values of both mispricing indicators, CMI_t^R (for the target asset) and $\text{CMI}_t^{R^*}$ (for the synthetic). The trading rule TR_t can take one of three values: $+1$, -1, or 0, indicating a "*long-short*", "*short-long*", or "*flat*" position, respectively. When no position is open (i.e., $TR_{t-1} = 0$), the rule opens a position only if there is simultaneous mispricing in both assets according to the thresholds D_l and D_u. Specifically,

- **Long target/Short synthetic (+1)**: Entered when both CMIs indicate the target asset is underpriced relative to the synthetic ($\text{CMI}_t^R \leq D_l$ **and** $\text{CMI}_t^{R^*} \geq D_u$).

- **Short target/Long synthetic (-1)**: Entered when both CMIs indicate the target asset is overpriced relative to the synthetic ($\text{CMI}_t^R \geq D_u$ **and** $\text{CMI}_t^{R^*} \leq D_l$).

Once a position is open (either $TR_{t-1} = +1$ or $TR_{t-1} = -1$), the logic checks each day whether the mispricing has corrected enough to trigger a take-profit condition or crossed critical boundaries that trigger a stop-loss. These checks apply to either of the two mispricing indices, so if correction or a stop-loss occurs in any one of them, the entire position is closed. Mathematically, this is captured by the "*OR*" clauses in the formula, which evaluate whether CMI_t^R or $\text{CMI}_t^{R^*}$ has crossed the zero line (for take-profit) or moved beyond the (S_l, S_u) band (for stop-loss). If one of these events occurs, then TR_t is set to 0, and the mispricing indices are both reset to zero for the next trading day. If neither a take-profit nor a stop-loss threshold is met, then the position remains unchanged, meaning TR_t simply inherits the previous value TR_{t-1}.

Intuitively, when both indicators are simultaneously misaligned (one significantly high and the other significantly low), the strategy deems it a strong signal to open a dollar-neutral position that is long the "*undervalued*" side and short the "*overvalued*" side. As soon as either index crosses back toward zero (suggesting partial correction of that asset's mispricing) or breaches a stop-loss boundary (indicating that the trade is moving unfavorably), the position is liquidated. This "*and-or*" logic helps filter out noise in the daily movements and more reliably captures episodes in which both assets appear to be drifting apart (opening a trade) and then swiftly catches at least one side reverting (closing the trade). We formally present this procedure in Algorithm 6.

As in Xie et al. (2016), we set $(D_l, D_u) = (-0.6, 0.6)$ and $(S_u, S_l) = (-2, 2)$ and we will explore other parametric choices in the robustness checks.

5. Empirical Application

5.1 Assumptions

To ensure transparency in our empirical analysis, we explicitly outline the critical assumptions underlying the implementation of our proposed pairs trading strategy. These assumptions reflect idealized market conditions necessary for theoretical feasibility and reproducibility of results.

Assumption 1 (Price Execution)**.** *All trades are executed at daily adjusted closing prices. This assumption requires sufficient market liquidity and depth to accommodate position entries and exits without significant price impact or execution delays.*

Assumption 2 (Short Selling Access)**.** *Unrestricted short selling is permitted for all assets, including the ability to maintain leveraged short positions. This encompasses having reliable access to securities lending facilities and the capacity to meet associated margin requirements.*

Assumption 3 (Leverage Capacity)**.** *Trading positions can employ substantial leverage on both long and short sides. This assumes access to margin facilities that permit position sizes meaningfully larger than the allocated capital base, subject to prevailing broker and regulatory requirements.*

While these assumptions may appear restrictive, recent developments in financial technology and market structure have made such trading conditions increasingly accessible. Modern electronic trading platforms like Alpaca, Interactive Brokers, and similar services now offer retail investors sophisticated capabilities previously reserved for institutional traders. These platforms provide programmatic trading interfaces, competitive margin rates, and extensive short-selling facilities that may align with our implementation requirements.

Cautionary note. *This paper is intended for academic purposes only and does not constitute financial advice. The strategies and methodologies discussed involve significant risks, including the potential loss of capital. Past performance is not indicative of future results, and the authors assume no liability for decisions made by individuals or entities based on the content of this research.*

References

C. Alexander. Optimal hedging using cointegration. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 357(1758):2039–2058, Aug. 1999. ISSN 1471-2962. doi: 10.1098/rsta.1999.0416. URL http://dx.doi.org/10.1098/rsta.1999.0416.

C. Alexander. *Market risk analysis, Practical Financial Econometrics*. John Wiley & Sons, 2008.

C. Alexander and A. Dimitriu. The cointegration alpha: Enhanced index tracking and long-short equity market neutral strategies. *SSRN Electronic Journal*, 2002. ISSN 1556-5068. doi: 10.2139/ssrn.315619. URL http://dx.doi.org/10.2139/ssrn.315619.

C. Alexander and A. Dimitriu. Indexing and statistical arbitrage. *The Journal of Portfolio Management*, 31(2):50–63, Jan. 2005a. ISSN 2168-8656. doi: 10.3905/jpm.2005.470578. URL http://dx.doi.org/10.3905/jpm.2005.470578.

C. Alexander and A. Dimitriu. Indexing, cointegration and equity market regimes. *International Journal of Finance & Economics*, 10(3):213–231, 2005b. ISSN 1099-1158. doi: 10.1002/ijfe.261. URL http://dx.doi.org/10.1002/ijfe.261.

D. A. Bowen and M. C. Hutchinson. Pairs trading in the uk equity market: risk and return. *The European Journal of Finance*, 22(14):1363–1387, Sept. 2014. ISSN 1466-4364. doi: 10.1080/1351847x.2014.953698. URL http://dx.doi.org/10.1080/1351847X.2014.953698.

R. Bradrania, D. Pirayesh Neghab, and M. Shafizadeh. State-dependent stock selection in index tracking: a machine learning approach. *Financial Markets and Portfolio Management*, 36(1): 1–28, Apr. 2021. ISSN 2373-8529. doi: 10.1007/s11408-021-00391-7. URL http://dx.doi.org/10.1007/s11408-021-00391-7.

J. Caldeira and G. V. Moura. Selection of a portfolio of pairs based on cointegration: A statistical arbitrage strategy. *SSRN Electronic Journal*, 2013. ISSN 1556-5068. doi: 10.2139/ssrn.2196391. URL http://dx.doi.org/10.2139/ssrn.2196391.

A. Cartea and S. Jaimungal. Algorithmic trading of co-integrated assets. *SSRN Electronic Journal*, 2015. ISSN 1556-5068. doi: 10.2139/ssrn.2637883. URL http://dx.doi.org/10.2139/ssrn.2637883.

H. J. Chen, S. J. Chen, Z. Chen, and F. Li. Empirical investigation of an equity pairs trading strategy. *Management Science*, 65(1):370–389, Jan. 2019. ISSN 1526-5501. doi: 10.1287/mnsc.2017.2825. URL http://dx.doi.org/10.1287/mnsc.2017.2825.

C. C. Chu and P. K. Chan. Mining profitable high frequency pairs trading forex signal using copula and deep neural network. In *2018 19th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, pages 312–316. IEEE, June 2018. doi: 10.1109/snpd.2018.8441125. URL http://dx.doi.org/10.1109/SNPD.2018.8441125.

B. Do and R. Faff. Does simple pairs trading still work? *Financial Analysts Journal*, 66(4):83–95, July 2010. ISSN 1938-3312. doi: 10.2469/faj.v66.n4.1. URL http://dx.doi.org/10.2469/faj.v66.n4.1.

B. Do, R. Faff, and K. Hamza. A new approach to modeling and estimation for pairs trading. In *Proceedings of 2006 financial management association European conference*, volume 1, pages 87–99. Citeseer, 2006.

R. J. Elliott, J. Van Der Hoek, and W. P. Malcolm. Pairs trading. *Quantitative Finance*, 5(3): 271–276, June 2005. ISSN 1469-7696. doi: 10.1080/14697680500149370. URL http://dx.doi.org/10.1080/14697680500149370.

E. Gatev, W. N. Goetzmann, and K. G. Rouwenhorst. Pairs trading: Performance of a relative-value arbitrage rule. *Review of Financial Studies*, 19(3):797–827, 2006. ISSN 1465-7368. doi: 10.1093/rfs/hhj020. URL http://dx.doi.org/10.1093/rfs/hhj020.

C. Han, Z. He, and A. J. W. Toh. Pairs trading via unsupervised learning. *European Journal of Operational Research*, 307(2):929–947, June 2023. ISSN 0377-2217. doi: 10.1016/j.ejor.2022.09.041. URL http://dx.doi.org/10.1016/j.ejor.2022.09.041.

F. He, A. Yarahmadi, and F. Soleymani. Investigation of multivariate pairs trading under copula approach with mixture distribution. *Applied Mathematics and Computation*, 472:128635, July 2024. ISSN 0096-3003. doi: 10.1016/j.amc.2024.128635. URL http://dx.doi.org/10.1016/j.amc.2024.128635.

N. Huck and K. Afawubo. Pairs trading and selection methods: is cointegration superior? *Applied Economics*, 47(6):599–613, Nov. 2014. ISSN 1466-4283. doi: 10.1080/00036846.2014.975417. URL http://dx.doi.org/10.1080/00036846.2014.975417.

K. Johansson, T. Schmelzer, and S. Boyd. Finding moving-band statistical arbitrages via convex-concave optimization. *Optimization and Engineering*, Oct. 2024. ISSN 1573-2924. doi: 10.1007/s11081-024-09933-0. URL http://dx.doi.org/10.1007/s11081-024-09933-0.

J. Joubert, O. Proskurin, I. Barziy, V. Pervushyna, H. Pei, and Y. Wang. *The Definitive Guide to Pairs Trading*, 2021. URL https://github.com/hudson-and-thames/definitive_guide_to_pairs_trading/blob/main/Definitive_Guide_to_Pairs_Trading.pdf.

C. Krauss. Statistical arbitrage pairs trading strategies: review and outlook. *Journal of Economic Surveys*, 31(2):513–545, May 2016. ISSN 1467-6419. doi: 10.1111/joes.12153. URL http://dx.doi.org/10.1111/joes.12153.

C. Krauss and J. Stübinger. Non-linear dependence modelling with bivariate copulas: statistical arbitrage pairs trading on the s&p 100. *Applied Economics*, 49(52):5352–5369, Apr. 2017. ISSN 1466-4283. doi: 10.1080/00036846.2017.1305097. URL http://dx.doi.org/10.1080/00036846.2017.1305097.

C. Lau, W. Xie, and Y. Wu. Multi-dimensional pairs trading using copulas. In *European Financial Management Association 2016 Annual Meetings June*, 2016.

R. Q. Liew and Y. Wu. Pairs trading: A copula approach. *Journal of Derivatives & Hedge Funds*, 19(1):12–30, Feb. 2013. ISSN 1753-965X. doi: 10.1057/jdhf.2013.1. URL http://dx.doi.org/10.1057/jdhf.2013.1.

P. S. Lintilhac and A. Tourin. Model-based pairs trading in the bitcoin markets. *Quantitative Finance*, 17(5):703–716, Nov. 2016. ISSN 1469-7696. doi: 10.1080/14697688.2016.1231928. URL http://dx.doi.org/10.1080/14697688.2016.1231928.

A. Min and C. Czado. Bayesian inference for multivariate copulas using pair-copula constructions. *Journal of Financial Econometrics*, 8(4):511–546, May 2010. ISSN 1479-8417. doi: 10.1093/jjfinec/nbp031. URL http://dx.doi.org/10.1093/jjfinec/nbp031.

K. Qureshi and T. Zaman. Pairs trading using a novel graphical matching approach. *arXiv preprint arXiv:2403.07998*, 2024.

H. Rad, R. K. Y. Low, and R. Faff. The profitability of pairs trading strategies: distance, cointegration and copula methods. *Quantitative Finance*, 16(10):1541–1558, Apr. 2016. ISSN 1469-7696. doi: 10.1080/14697688.2016.1164337. URL http://dx.doi.org/10.1080/14697688.2016.1164337.

F. Rotondi and F. Russo. Machine learning for pairs trading: a clustering-based approach. 2025. doi: 10.2139/ssrn.5080998. URL http://dx.doi.org/10.2139/ssrn.5080998.

R. Roychoudhury, R. Bhagtani, and A. Daftari. Pairs trading using clustering and deep reinforcement learning. *SSRN Electronic Journal*, 2023. ISSN 1556-5068. doi: 10.2139/ssrn.4504599. URL http://dx.doi.org/10.2139/ssrn.4504599.

F. A. Sabino da Silva, F. A. Ziegelmann, and J. F. Caldeira. A pairs trading strategy based on mixed copulas. *The Quarterly Review of Economics and Finance*, 87:16–34, Feb. 2023. ISSN 1062-9769. doi: 10.1016/j.qref.2022.10.007. URL http://dx.doi.org/10.1016/j.qref.2022.10.007.

S. M. Sarmento and N. Horta. Enhancing a pairs trading strategy with the application of machine learning. *Expert Systems with Applications*, 158:113490, Nov. 2020. ISSN 0957-4174. doi: 10.1016/j.eswa.2020.113490. URL http://dx.doi.org/10.1016/j.eswa.2020.113490.

L. Shu, F. Shi, and G. Tian. High-dimensional index tracking based on the adaptive elastic net. *Quantitative Finance*, 20(9):1513–1530, Apr. 2020. ISSN 1469-7696. doi: 10.1080/14697688.2020.1737328. URL http://dx.doi.org/10.1080/14697688.2020.1737328.

Y. Stander, D. Marais, and I. Botha. Trading strategies with copulas. *Journal of Economic and Financial Sciences*, 6(1):83–107, 2013.

M. Tadi and J. Witzany. Copula-based trading of cointegrated cryptocurrency pairs. *Financial Innovation*, 11(1), Jan. 2025. ISSN 2199-4730. doi: 10.1186/s40854-024-00702-7. URL http://dx.doi.org/10.1186/s40854-024-00702-7.

G. Vidyamurthy. Pairs trading: Quantitative methods and analysis, 2004.

P. Wang and X. Ding. Pairs trading strategy based on copula-garch model. In *4TH International Scientific Conference of Alkafeel University (ISCKU 2022)*, volume 2977, page 080001. AIP Publishing, 2023. doi: 10.1063/5.0181010. URL http://dx.doi.org/10.1063/5.0181010.

W. Xie, R. Q. Liew, Y. Wu, and X. Zou. Pairs trading with copulas. *The Journal of Trading*, 11 (3):41–52, June 2016. ISSN 2168-8427. doi: 10.3905/jot.2016.11.3.041. URL http://dx.doi.org/10.3905/jot.2016.11.3.041.

Z. Zeng and C.-G. Lee. Pairs trading: optimal thresholds and profitability. *Quantitative Finance*, 14(11):1881–1893, Oct. 2014. ISSN 1469-7696. doi: 10.1080/14697688.2014.917806. URL http://dx.doi.org/10.1080/14697688.2014.917806.

T. Z. Zhi, X. Wenjun, W. Yuan, and X. Liming. Dynamic copula framework for pairs trading. Technical report, Working Paper, 2017.

A. Online Appendix

A.1 Algorithms

Algorithm 1. Sparse Synthetic Control

```
Require:
  Target asset log-prices $\mathbf{y} = [y_t]_{t=1}^{T} \in \mathbb{R}^T$
  Donor pool log-prices $\mathbf{X} = [x_{1t}, ..., x_{Nt}]_{t=1}^{T} \in \mathbb{R}^{T \times N}$
  Maximum number of assets $K \in \mathbb{N}$ with $K \leq N$
Ensure: Sparse weight vector $\mathbf{w}^* \in \mathbb{R}^N$
function SyntheticControl($\mathbf{y}, \mathbf{X}, K$)
    # Stage 1: Unrestricted optimization
    $\mathbf{w}^{(1)} = \arg\min_{\mathbf{w} \in \mathbb{R}^N} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$ s.t. $\mathbf{1}^\top \mathbf{w} = 1$    ▷ Solve full least squares problem
    # Stage 2: Support selection
    $\mathcal{I} \leftarrow \{i : |w_i^{(1)}|$ among $K$ largest$\}$    ▷ Select $K$ largest weights
    $\mathbf{X}_{\mathcal{I}} \leftarrow [\mathbf{x}_{\mathcal{I}_1}, \dots, \mathbf{x}_{\mathcal{I}_K}]$    ▷ Restricted donor matrix
    # Stage 3: Restricted optimization
    $\mathbf{w}^{(2)} = \arg\min_{\mathbf{w}_{\mathcal{I}} \in \mathbb{R}^K} \|\mathbf{y} - \mathbf{X}_{\mathcal{I}}\mathbf{w}_{\mathcal{I}}\|_2^2$ s.t. $\mathbf{1}^\top \mathbf{w}_{\mathcal{I}} = 1$    ▷ Solve restricted program
    for each $i \in \{1, \dots, N\}$ do
        $w_i^* \leftarrow w_j^{(2)}$ if $i = \mathcal{I}_j$, else 0    ▷ Construct full weights
    end for
    return $\mathbf{w}^*$
end function
```

Algorithm 2. Copula Fitting

```
Require:
Target returns $\mathbf{r} = [r_t]_{t=2}^{T} \in \mathbb{R}^{T-1}$
Synthetic returns $\mathbf{r}^* = [r_t^*]_{t=2}^{T} \in \mathbb{R}^{T-1}$
Parametric copula families $\mathcal{C} = \{C_\theta : \theta \in \Theta\}$
Numerical tolerance $\epsilon = 10^{-5}$
Ensure: Marginal ECDFs $\hat{F}_R, \hat{F}_{R^*}$ and fitted copulas $\{C_{\hat{\theta}}\}_{C_\theta \in \mathcal{C}}$
function CopulaFit($\mathbf{r}, \mathbf{r}^*$)
    # Construct linearly interpolated ECDFs
    for each return series $\mathbf{x} \in \{\mathbf{r}, \mathbf{r}^*\}$ do
        Sort unique values: $x_{(1)} < \cdots < x_{(m)}$
        $p_i \leftarrow \frac{1}{T-1} \sum_{t=2}^{T} \mathbb{I}(x_t \leq x_{(i)})$    ▷ Compute empirical probabilities
        $\hat{F}_X(x) \leftarrow p_i + (p_{i+1} - p_i) \frac{x - x_{(i)}}{x_{(i+1)} - x_{(i)}}$ for $x \in [x_{(i)}, x_{(i+1)}]$:    ▷ Piecewise linear interpolation
    end for

    # Apply probability integral transform
    for $t \in \{2, \ldots, T\}$ do
        $u_t \leftarrow \max\{\epsilon, \min\{\hat{F}_R(r_t), 1 - \epsilon\}\}$    ▷ Adjust ECDF outputs to tolerance level $\epsilon$
        $v_t \leftarrow \max\{\epsilon, \min\{\hat{F}_{R^*}(r_t^*), 1 - \epsilon\}\}$
    end for

    # Fit each copula family
    for each copula family $C_\theta \in \mathcal{C}$ do
        $\hat{\theta} \leftarrow \arg\max_{\theta \in \Theta} \sum_{t=2}^{T} \ln c_\theta(u_t, v_t)$    ▷ Estimate parameters via maximum likelihood
        $\ell(\hat{\theta}) \leftarrow \sum_{t=2}^{T} \ln c_{\hat{\theta}}(u_t, v_t)$    ▷ Obtain maximized likelihood
    end for

    return $\hat{F}_R, \hat{F}_{R^*}, \{C_{\hat{\theta}}\}_{C_\theta \in \mathcal{C}}$
end function
```

Algorithm 3. Mispricing Indices Calculation

```
Require:
    Target return $r_t$, synthetic return $r_t^*$
    Optimal copula $C_{\hat{\theta}}$
    Marginal ECDFs $\hat{F}_R, \hat{F}_{R^*}$
Ensure: Mispricing indices $MI_t^{R|R^*}, MI_t^{R^*|R}$
function MispricingIndices($r_t, r_t^*, C_{\hat{\theta}}, \hat{F}_R, \hat{F}_{R^*}$)
    $u_t \leftarrow \hat{F}_R(r_t)$, $v_t \leftarrow \hat{F}_{R^*}(r_t^*)$    ▷ Compute uniform marginals (pseudo-observations)
    $MI_t^{R|R^*} \leftarrow \frac{\partial C_{\hat{\theta}}(u_t, v_t)}{\partial v_t}$    ▷ Compute target-synthetic MI
    $MI_t^{R^*|R} \leftarrow \frac{\partial C_{\hat{\theta}}(u_t, v_t)}{\partial u_t}$    ▷ Compute synthetic-target MI
    return $MI_t^{R|R^*}, MI_t^{R^*|R}$
end function
```

Algorithm 4. Update Cumulative Mispricing Index (CMI)

```
Require:
    Mispricing indices: $(\mathrm{MI}_t^{R|R^*}, \mathrm{MI}_t^{R^*|R})$
    Previous CMIs: $(\mathrm{CMI}_{t-1}^{R}, \mathrm{CMI}_{t-1}^{R^*})$
    Reset flag: reset
Ensure: Updated CMIs: $(\mathrm{CMI}_t^{R}, \mathrm{CMI}_t^{R^*})$
function UpdateCMI($\mathrm{MI}_t^{R|R^*}, \mathrm{MI}_t^{R^*|R}, \mathrm{CMI}_{t-1}^{R}, \mathrm{CMI}_{t-1}^{R^*}$, reset)
    if reset then
        $\mathrm{CMI}_t^{R} \leftarrow 0$, $\mathrm{CMI}_t^{R^*} \leftarrow 0$    ▷ Reset the CMIs to 0
    else
        $\mathrm{CMI}_t^{R} \leftarrow \mathrm{CMI}_{t-1}^{R} + (\mathrm{MI}_t^{R|R^*} - 0.5)$    ▷ Update target CMIs with new realization of MI
        $\mathrm{CMI}_t^{R^*} \leftarrow \mathrm{CMI}_{t-1}^{R^*} + (\mathrm{MI}_t^{R^*|R} - 0.5)$
    end if
    return $(\mathrm{CMI}_t^{R}, \mathrm{CMI}_t^{R^*})$
end function
```

Algorithm 5. Trading Rule

```
Require: Mispricing indices $\mathrm{CMI}_t^R, \mathrm{CMI}_t^{R^*}$ and thresholds $D_l, D_u, S_l, S_u$
Ensure: Trading position $TR_t \in \{-1, 0, +1\}$

function TradingRule( $\mathrm{CMI}_t^R, \mathrm{CMI}_t^{R^*},\ D_l,\ D_u,\ S_l,\ S_u$ )
    if $TR_{t-1} = 0$ then                                         ▷ No existing position
        if $\mathrm{CMI}_t^R \le D_l$ and $\mathrm{CMI}_t^{R^*} \ge D_u$ then
            $TR_t \leftarrow +1$                                   ▷ Long target, short synthetic
        else if $\mathrm{CMI}_t^R \ge D_u$ and $\mathrm{CMI}_t^{R^*} \le D_l$ then
            $TR_t \leftarrow -1$                                   ▷ Short target, long synthetic
        else
            $TR_t \leftarrow 0$                                    ▷ Remain flat
        end if
    else if $TR_{t-1} = +1$ then                                   ▷ Currently long target, short synthetic
        if ($\mathrm{CMI}_t^R \ge 0$ or $\mathrm{CMI}_t^{R^*} \le 0$) or ($\mathrm{CMI}_t^R \le S_l$ or $\mathrm{CMI}_t^{R^*} \ge S_u$) then
            $TR_t \leftarrow 0$                                    ▷ Close position (take profit or stop-loss)
            Reset $\mathrm{CMI}_t^R \leftarrow 0$ and $\mathrm{CMI}_t^{R^*} \leftarrow 0$
        else
            $TR_t \leftarrow +1$                                   ▷ Maintain current position
        end if
    else if $TR_{t-1} = -1$ then                                   ▷ Currently short target, long synthetic
        if ($\mathrm{CMI}_t^R \le 0$ or $\mathrm{CMI}_t^{R^*} \ge 0$) or ($\mathrm{CMI}_t^R \ge S_u$ or $\mathrm{CMI}_t^{R^*} \le S_l$) then
            $TR_t \leftarrow 0$                                    ▷ Close position (take profit or stop-loss)
            Reset $\mathrm{CMI}_t^R \leftarrow 0$ and $\mathrm{CMI}_t^{R^*} \leftarrow 0$
        else
            $TR_t \leftarrow -1$                                   ▷ Maintain current position
        end if
    end if
    return $TR_t$
end function
```

Algorithm 6. Main. "*Pairs-trading a Sparse Synthetic Control*"

```
Require:
    Target asset log-prices $\mathbf{y} = [y_t]_{t=1}^T$
    Donor pool log-prices $\mathbf{X} = [x_{1t}, ..., x_{Nt}]_{t=1}^T$
    Maximum number of assets $K \in \mathbb{N}$ with $K \leq N$
    Entry thresholds $(D_l, D_u)$, stop-loss thresholds $(S_l, S_u)$
    Parametric copula families $\mathcal{C} = \{C_\theta : \theta \in \Theta\}$
Ensure: Trading signals $\{TR_t\}_{t=1}^T$

procedure Main($\mathbf{y}, \mathbf{X}, K, D_l, D_u, S_l, S_u, \mathcal{C}$)
    $\mathbf{w}^* \leftarrow$ SyntheticControl($\mathbf{y}, \mathbf{X}, K$)          ▷ Construct synthetic asset
    $\mathbf{y}^* \leftarrow \mathbf{X}\mathbf{w}^*$
    $\mathbf{r} \leftarrow \text{diff}(\mathbf{y})$, $\mathbf{r}^* \leftarrow \text{diff}(\mathbf{y}^*)$          ▷ Compute returns
    $C_{\hat{\theta}}, \hat{F}_R, \hat{F}_{R^*} \leftarrow$ CopulaFit($\mathbf{r}, \mathbf{r}^*$)          ▷ Fit copula
    Initialize $TR_0 \leftarrow 0$, $\text{CMI}_0^R \leftarrow 0$, $\text{CMI}_0^{R^*} \leftarrow 0$
    for $t = 1$ to $T$ do
        $\text{MI}_t^{R|R^*}, \text{MI}_t^{R^*|R} \leftarrow$ MispricingIndices($r_t, r_t^*, C_{\hat{\theta}}, \hat{F}_R, \hat{F}_{R^*}$)
        $TR_t \leftarrow$ TradingRule($\text{CMI}_{t-1}^R, \text{CMI}_{t-1}^{R^*}, TR_{t-1}, D_l, D_u, S_l, S_u$)
        reset $\leftarrow (TR_t = 0$ and $TR_{t-1} \neq 0)$          ▷ Reset CMI if position closed
        $\text{CMI}_t^R, \text{CMI}_t^{R^*} \leftarrow$ UpdateCMI($MI_t^{R|R^*}, \text{MI}_t^{R^*|R}, \text{CMI}_{t-1}^R, \text{CMI}_{t-1}^{R^*}$, reset)
    end for
    return $\{TR_t\}_{t=1}^T$
end procedure
```

A.2 Cointegration Meets Synthetic Controls: A Formal Equivalence

In this appendix section, we develop a formal argument showing how, under some stringent assumptions, our notion of *synthetic control* can be viewed as a special case of *cointegration.* This connection underlies the intuition that, when one normalizes the first variable of a cointegrated system to 1, the remaining cointegration relationships effectively produce the *synthetic* version of the first variable when the cointegration vector satisfies a specific restriction.

Let $\{y_{i,t}\}_{t=1}^T$ denote the time series sequence of log-prices for each asset $i \in \{1, \dots, N\}$. Throughout, we assume each $y_{i,t}$ is an $I(1)$ process (integrated of order 1). Formally, an $I(1)$ process is one that becomes *stationary* (and typically ergodic) upon differencing once: $\Delta y_{i,t} := y_{i,t} - y_{i,t-1} \sim I(0)$. The notion of cointegration, due to Engle and Granger, is central in analyzing potentially long-run equilibria among these variables.

Definition 2 (Engle and Granger (1987)). *The components of $\mathbf{y}_t := [y_{1t}, ..., y_{Nt}]$ are said to be cointegrated of order d, b, denoted $\mathbf{y}_t \sim CI(d, b)$, if (a) all components of $\mathbf{y}_t$ are $I(d)$ and (b) a vector $\boldsymbol{\beta} \neq 0$ exists so that $\boldsymbol{\beta}'\mathbf{y}_t \sim I(d-b)$, $b > 0$. The vector $\boldsymbol{\beta}$ is called the cointegrating vector.*

Definition 3 (Synthetic Control). *Let $\{y_1, y_2, \dots, y_n\}$ be a collection of random variables, where y_1 is the "target" variable and $\mathbf{y}_{2:n} = (y_2, \dots, y_n)$ constitute the "donor pool". A* synthetic control *for y_1 is constructed by choosing weights $\mathbf{w}$ in the $(n-1)$-dimensional space $\mathcal{W} := \{\mathbf{w} \in \mathbb{R}^{n-1}_{+} : \sum_{j=2}^{n} w_j = 1\}$ that satisfy $\mathbf{w} = \arg\min_{w \in \mathcal{W}} \sum_{t=1}^{T}(y_{1,t} - \mathbf{w}'\mathbf{y}_{2:n,t})^2$.*

Given that cointegration relationships prevail up to scale and sign changes, then, under suitable conditions on the cointegration vector, there exists a nontrivial constant κ that allows us to reinterpret the cointegration relationship as one of a synthetic control. In particular,

Proposition 1. *For a cointegrated vector $\mathbf{y}$ with rank r, if (at least) one of the cointegrating vectors $\boldsymbol{\beta}$ satisfies the restriction $\mathcal{R} = \{\mathbf{1}'\boldsymbol{\beta} = 0\}$, then we can scale the cointegration vector by $\kappa = 1/\beta_i$ such that $\kappa\boldsymbol{\beta}'\mathbf{y}$ is stationary and describes a "synthetic control" relationship (as per Definition 3) between y_i and $\mathbf{y}_{-i}$.*

Proof. The proof is straightforward. For a cointegration vector $\boldsymbol{\beta}$ where $\mathcal{R}$ holds, we have that $\mathbf{1}'\boldsymbol{\beta} = \sum_{j=1}^{n} \beta_j = 0$, which trivially implies $\beta_i = -\sum_{j \neq i} \beta_j$. For the sake of the proof, set that β_i to the first component (β_1). Then $\beta_1 = -\sum_{j=2}^{n} \beta_j$ and $\kappa = (\beta_1)^{-1} = -(\sum_{j=2}^{n} \beta_j)^{-1}$

$$\kappa\boldsymbol{\beta}'\mathbf{y} = \frac{1}{\beta_1}[\beta_1 \;\; \boldsymbol{\beta}_{2:n}]\mathbf{y}_t = \begin{bmatrix} 1 & \frac{-\boldsymbol{\beta}'_{2:n}}{\sum_{j=2}^{n} \beta_j} \end{bmatrix} \begin{bmatrix} y_1 \\ \mathbf{y}_{2:n} \end{bmatrix} = y_1 - \frac{\beta_2}{\sum_{j=2}^{n} \beta_j} y_2 - \dots - \frac{\beta_n}{\sum_{j=2}^{n} \beta_j} y_n \sim I(0)$$

describes a stationary cointegration relationship in $\mathbf{y}$, and since

$$\begin{aligned} y_1 &= \frac{\beta_2}{\sum_{j=2}^{n} \beta_j} y_2 + \dots + \frac{\beta_n}{\sum_{j=2}^{n} \beta_j} y_n + \epsilon \\ &= \mathbf{w}'\mathbf{y}_{2:n} + \epsilon \end{aligned}$$

with $\epsilon \sim I(0)$ and $\mathbf{w} := \left(\frac{\beta_2}{\sum_{j=2}^{n} \beta_j}, ..., \frac{\beta_n}{\sum_{j=2}^{n} \beta_j}\right)' \in \mathcal{W}$, then this relationship is endowed with a synthetic control structure. A similar reasoning applies to any other β_i different from β_1. □