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: 26th 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 an ℓ_1-regularized least squares optimization, which automatically selects a sparse subset of assets from a broad donor pool while maintaining 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, Copula, Basket trading

* 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.

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 is structured as follows. We begin in section 1.1 with a comprehensive review of the relevant literature, exploring pairs trading, synthetic control methods, and copula-based dependence modeling. section 2 presents our methodological framework, detailing how we employ sparse synthetic control and copula families to construct a robust trading signal. In section 3, we introduce the mispricing index (MI) strategy, adapted to incorporate copula-driven signals, and define the cumulative mispricing index (CMI) as a key component of our trading approach. Section 4 shows the results of our trading strategy, and we conclude in section 5 by synthesizing key insights and proposing future research directions.

1.1 Literature Review

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 using a sparse 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. Then, we can build a synthetic asset time series $\mathbf{y}^*$ as

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

where the weights $\mathbf{w}^* = [w_1^*, ..., w_N^*]^\top \in \mathbb{R}^N$ are determined via an ℓ_1-regularized least squares optimization problem

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

where $\|\mathbf{w}\|_1 = \sum_{i=1}^N |w_i|$ denotes the ℓ_1-norm of the weight vector and $\lambda > 0$ is a regularization parameter that controls the level of sparsity. The ℓ_1 regularization, also known as the lasso penalty, induces sparsity by shrinking some weights exactly to zero, effectively performing feature selection

among the donor assets. This sparsity-inducing property stems from the non-differentiability of the penalty term at the origin. The practical implementation of this procedure is given in Algorithm 1.

The optimization problem possesses several key theoretical and practical features that make it particularly suitable for our application. First, the combination of a quadratic loss function with the convex ℓ_1-penalty and affine constraint guarantees a unique solution under mild regularity conditions. Second, the regularization parameter λ (optimally selected through cross-validation) provides direct control over the sparsity level, with larger values yielding solutions with fewer non-zero weights. Third, the simplex constraint $\mathbf{1}^\top \mathbf{w} = 1$ ensures interpretability of the synthetic control as a weighted portfolio of donor assets. We don't impose a convex hull restriction, which effectively means that we allow for negative ways in the synthetic asset.

The resulting weight vector $\mathbf{w}^*$ will typically have many components equal or very close to zero, with the number of non-zero weights decreasing as λ increases. In practice, we identify the support of non-zero weights through thresholding:

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

where $\epsilon > 0$ is a small tolerance threshold (in our application, $\epsilon \approx 10^{-5}$). The final synthetic asset is then constructed using only assets in $\mathcal{I}$. We chose to sparsify the synthetic control using lasso instead of solving a cardinality-constrained program as the former is able to maintain sparse exposures while enjoying vast computational advantages. Moreover, the convex nature of the problem permits efficient solution via proximal algorithms or quadratic programming techniques, making it suitable for high-dimensional donor pools. For a detailed discussion, see Appendix Section A.2.

The reader may draw similarities of this process with the Engle-Granger procedure for estimating the cointegration vector associated to the target asset and the donor pool. If we don't impose the ℓ_1-penalty, it can be shown that, under some conditions, the procedure of finding the weights of the synthetic control is equivalent to finding the cointegration vector. For a formal discussion see Appendix Section A.1.

To evaluate our methodology empirically, we implement the synthetic control approach using daily adjusted-close price data from S&P500 constituents. We select NVIDIA (NVDA) as our target asset and construct the donor pool from the remaining index components. The full sample runs from January 2010 to January 2025 and is partitioned chronologically, with 70% allocated to the training period for model estimation and the remaining 30% reserved for out-of-sample testing. The optimization procedure detailed in the previous section yields the optimal weights presented in Table 1. The resulting synthetic control comprises 27 stocks with non-zero weights that sum to unity, distributed between long positions (positive weights) and short positions (negative

weights). This sparse portfolio structure effectively defines a tradeable basket that can be executed simultaneously through standard ETF-like basket trading mechanisms.

The time series evolution of the target and synthetic log-prices is shown in Figure 1. As we can see, the fit is really good in the training sample, but as we move out of sample, the spread between the two series becomes more volatile and the two series seem to diverge in recent years. As we will see later, we should not worry too much about the log-price fit, as the trading strategy capitalizes on the mean-reversion of returns (rather than log-prices).

[Insert Figure 1 about here]

2.2 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, and (3) selection of an appropriate copula family

2.2.1 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.

Figure 2 presents a comparative analysis of the empirical cumulative distribution functions (ECDFs) between the target and synthetic assets across both training and testing periods, revealing several notable patterns. The analysis of ECDFs in both return and price spaces provides crucial insights into the effectiveness of our synthetic replication strategy. In the training period (top panels), we observe an almost perfect alignment between the target (red) and synthetic (blue) ECDFs in both return and price spaces. The log-returns distributions (top-left) exhibit the characteristic S-shaped curve typical of financial returns, with steep slopes around zero and thinner tails, suggesting that both assets capture similar distributional properties including volatility patterns and extreme events. The log-price ECDFs (top-right) show an equally strong correspondence, indicating successful replication of the price level dynamics during the training phase. However, the testing period (bottom panels) reveals a striking contrast between returns and prices that has important implications for our trading strategy. The log-returns ECDFs (bottom-left) maintain their strong alignment, with both target and synthetic assets continuing to share nearly identical distributional properties. This persistence in the return space is particularly noteworthy as it suggests that our synthetic construction preserves its ability to mimic the target's return dynamics even out-of-sample. In contrast, the log-price ECDFs (bottom-right) exhibit notable divergence

in the testing period, especially in the upper quantiles (above the 0.6 probability level). This divergence manifests as a systematic deviation between the target and synthetic price paths, with the synthetic asset generally showing lower price levels than the target for given probability levels. This pattern suggests that while the return dynamics remain well-matched, price levels can drift apart over time due to the cumulative effect of small return differences. This asymmetric behavior between returns and prices provides strong empirical justification for our copula-based approach, which focuses on modeling return dependencies rather than price-level relationships. The stability of the return distributions, even when price levels diverge, indicates that return-based trading signals may offer more reliable indicators of relative mispricings than traditional price-level methods. Furthermore, the observed price divergence in the testing period actually represents potential trading opportunities, as it suggests periods where the target and synthetic assets have moved out of their historical relationship, potentially creating conditions for profitable mean reversion trades.

[Insert Figure 2 about here]

Figure 3 reveals several important patterns in the relationship between target and synthetic assets across different dimensions. In the returns series (left panels), we observe considerable dispersion around the 45-degree line in both training and testing periods, with points forming a cloud-like pattern that suggests a consistent dependency structure. This dispersion reflects the inherent volatility in daily returns, yet the overall symmetric pattern around the diagonal indicates that the synthetic asset effectively captures the statistical properties of the target's return distribution.

The price series (right panels) exhibits markedly different behavior. During the training period, we observe a tight clustering of points along the diagonal, indicating that the synthetic asset closely tracks the price level distribution of the target. However, the testing period reveals a subtle but important deviation from this pattern, with increased dispersion and slight systematic deviations from the 45-degree line, particularly in the middle range of the distribution (0.4-0.6 probability region). This divergence suggests that while the synthetic asset maintains similar distributional properties in returns space, the accumulated price levels begin to show some drift in the out-of-sample period.

The contrast between return and price space behaviors is particularly instructive. The relatively stable pattern in returns space, even during the testing period, suggests that the copula-based approach successfully captures the underlying return dynamics. Meanwhile, the gradual price divergence in the testing period highlights the challenges of maintaining long-term price level correspondence between target and synthetic assets, a common phenomenon in pairs trading that often creates trading opportunities.

This dual perspective-stability in returns but drift in prices-provides empirical support for our strategy's focus on return-based copula modeling rather than price-level relationships, as it appears to offer more robust statistical properties for trading signal generation.

[Insert Figure 3 about here]

2.2.2 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 (a formal description of the copula fitting procedure can be found in Algorithm 2.):

- **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 $\Theta \subset \mathbb{R}^2_+$

Figure 4 visualizes the dependence structures of different parametric copula families through density heatmaps generated from samples. The Archimedean copulas each exhibit distinct characteristics: Clayton demonstrates strong lower tail dependence, Gumbel and Joe show pronounced upper tail dependence, while Frank captures dependence concentrated in both tails. Among elliptical copulas, the Gaussian shows symmetric dependence with moderate tails, while the Student-t accommodates heavier tails in both directions. The N14, as a Clayton-Gumbel mixture, combines properties of both families to achieve a more flexible dependence structure. These varying patterns underscore how each family could potentially capture different aspects of movement between the target and synthetic asset.

[Insert Figure 4 about here]

2.2.3 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:

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

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.

[Insert Table 2 about here]

Table 2 presents the fitting results for different copula families. The comparative analysis of copula specifications reveals a clear dominance of elliptical copulas in modeling the dependence structure between our target and synthetic returns. The Student-t copula achieves the best fit across all information criteria, followed closely by the N14 mixed copula and the Gaussian copula, with substantially better performance than all Archimedean specifications. This hierarchy in goodness-of-fit measures, particularly the superior performance of symmetric specifications (Student-t, Gaussian) over asymmetric ones (Clayton, Gumbel, Joe), indicates that the dependence structure is predominantly symmetric, which aligns with the theoretical construction of the synthetic control as a tracking portfolio.

The significant improvement in fit from Gaussian to Student-t copula (ΔSIC $\approx$ 90 points) demonstrates that while symmetry is essential, the joint distribution exhibits heavier tails than implied by normal dependence. This finding is further supported by the Frank copula's performance, which achieves better fit than other Archimedean copulas but worse than elliptical specifications, suggesting that symmetric dependence without tail dependence is insufficient for capturing the full relationship. The relatively poor fit of copulas with exclusive lower tail (Clayton, SIC: -1168.92) or upper tail (Joe, SIC: -671.50) dependence reinforces that the target-synthetic relationship maintains its symmetric structure even during extreme market movements, a desirable property for the stability of our trading strategy.

3. 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.

3.1 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.

3.2 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:

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

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

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 sensitivity to other parametric choices in future robustness checks.

4. Results

Our pairs trading strategy represents a specialized implementation of basket trading, where a synthetic asset is constructed through an optimally weighted basket of securities. The strategy's feasibility rests on several key operational assumptions and market structure considerations that warrant careful examination. The basket's composition demonstrates sophisticated risk distribution across multiple market segments, with our synthetic control comprising 27 assets spanning diverse sectors. Significant long positions in AME (41.08%), LUV (33.31%), and TFC (25.60%) are counterbalanced by strategic shorts in ADSK (-42.25%) and UNP (-25.77%), creating a well-diversified exposure structure that provides distinct advantages over traditional single-stock pairs.

This sophisticated composition yields multiple advantages through its sector diversity. By spanning financials, technology, transportation, and other sectors, our approach substantially reduces idiosyncratic risk compared to single-stock pairs, effectively mitigating the impact of company-specific events or sector-wide shocks that typically destabilize simpler pairs trading arrangements. The implementation efficiency is further enhanced by modern execution systems that can treat these 27 components as a single basket order, substantially reducing transaction costs and bid-ask

spread impact through optimized order routing.

The practical implementation of this strategy requires careful consideration of several critical operational requirements. Most importantly, the approach necessitates access to sophisticated basket trading capabilities that can handle simultaneous execution of multiple components–a requirement readily met by major institutional brokers offering advanced smart order routing services. To ensure consistent execution quality, all basket components must maintain sufficient liquidity; we addressed this constraint by restricting our universe to S&P 500 constituents, thereby guaranteeing deep and reliable trading volumes across all components.

The strategy's effectiveness ultimately depends on maintaining strict dollar neutrality through equal but opposite positions in the target and synthetic basket. This market-neutral construction demands precise execution coordination across all components, a challenge effectively addressed by modern electronic trading systems capable of processing complex basket orders as single units.

Figure 5 illustrates the evolution of position sizes for both target and synthetic assets across different copula specifications over our out-of-sample period from July 2020 to January 2025. For each copula family, the strategy maintains dollar-neutral positions by simultaneously taking offsetting long and short positions in the target (red) and synthetic (blue) assets. Position sizes (in shares) are normalized relative to an initial equity of \$1, with values between -0.2 and 0.2 representing the number of shares allocated to each side of the trade. The parallel movements in the red and blue lines across all panels reflect the strategy's market-neutral construction, where positions are always equal in magnitude but opposite in direction.

[Insert Figure 5 about here]

Table 3 reveals the out-of-sample performance across copula specifications. The N14 mixed copula generates the highest total return (77.82%) and annualized return (17.26%), while maintaining moderate volatility (4.35%). This superior performance is reflected in its high Sharpe (3.97) and Sortino (5.75) ratios, matching the Clayton copula's risk-adjusted metrics but with better absolute returns. The Gaussian copula, despite showing the most conservative returns (62.70%), still achieves respectable risk-adjusted performance with a Sharpe ratio of 3.14, though it lags behind all other specifications. A particularly noteworthy pattern emerges in the downside risk metrics. The Frank copula, while generating modest returns (66.53%), exhibits the lowest volatility (3.97%) and maintains a remarkably low maximum drawdown (1.36%), suggesting it effectively filters out noisy trading signals. In contrast, the Joe copula shows the highest maximum drawdown (2.57%) despite having the highest win rate (36.62%), indicating that its emphasis on upper tail dependence may lead to larger adverse price movements before positions converge. The Student-t copula

strikes a balance between these extremes, delivering strong total returns (74.63%) with moderate maximum drawdown (2.15%) and a competitive Sharpe ratio (3.60), demonstrating that its symmetric tail dependence effectively captures the risk-return tradeoff inherent in our pairs trading strategy.

[Insert Table 3 about here]

Figure 6 displays the cumulative returns for our pairs trading strategy across different copula specifications. The performance hierarchy is well-defined throughout the out-of-sample period: the N14 mixed copula consistently outperforms all other specifications, achieving approximately 78% cumulative return. A second group comprising the Student-t, Clayton, and Gumbel copulas tracks closely together, delivering returns around 73-75%. Joe and Frank copulas form a third tier with returns near 67%, while the Gaussian copula lags notably behind all other specifications with about 63% cumulative return. This ordering largely aligns with the sophistication of tail dependence modeling in each specification, suggesting that more flexible approaches to capturing the joint distribution of target and synthetic returns translate directly into superior trading performance.

[Insert Figure 6 about here]

Risk-adjusted returns For each copula model, we run the regression

$$\mathcal{R}_t^c = \alpha + \boldsymbol{\beta}' \mathbf{f}_t + \epsilon_t$$

where $\mathcal{R}_t^c$ are the excess returns of our pairs trading strategy for copula family c, and $\mathbf{f}_t$ represents a particular combination of the Fama-French factors: $\texttt{MKT\_RF}_t$, $\texttt{SMB}_t$, $\texttt{HML}_t$, $\texttt{RMW}_t$, $\texttt{CMA}_t$, $\texttt{MOM}_t$, $\texttt{ST\_REV}_t$, $\texttt{LT\_REV}_t$. Since we consider combinations of 8 factors, we run $2^8 = 256$ different regressions for each copula. This delivers a positive significant α of 0.0004 – 0.0006 for all regressions specifications across all copula models, which is equivalent to an annualized α of 10.08% – 15.12%. The significance of risk-factors varies through copula models. In particular:

- Gumbel copula shows limited factor exposure, with only the Size (SMB) factor showing statistical significance and the Value (HML) factor showing mild significance. The Short-term Reversal (ST_REV) factor exhibits mild significance, although the overall factor model reliability is questionable.
- Frank copula returns demonstrate significant exposure primarily to the Short-term Reversal (ST_REV) factor, with other factors showing no consistent statistical significance.
- Clayton copula strategy returns are significantly exposed to market risk, with MKT_RF being the sole statistically significant factor across all model specifications.

- Joe copula exhibits mild sensitivity to market risk, with MKT_RF showing moderate significance. The single-factor model incorporating only MKT_RF emerges as the most statistically significant specification.
- N14 copula returns show significant exposure to the market factor. The model's explanatory power improves notably when incorporating the Short-term Reversal factor.
- Gaussian copula demonstrates significant exposure to market risk, with the Short-term Reversal factor showing strong statistical significance (significant at the 1% level).
- Student-t copula returns exhibit mild sensitivity to the Profitability factor (RMW), suggesting the potential relevance of a single-factor model focused solely on RMW.

5. Conclusions

This paper introduces a novel pairs trading framework that integrates sparse synthetic control methods with copula-based dependence modeling. Our methodology addresses several key limitations of traditional pairs trading approaches by combining adaptive synthetic asset construction with flexible dependence modeling. The empirical results demonstrate the effectiveness of this integrated approach across multiple dimensions.

The sparse synthetic control methodology successfully constructs parsimonious tracking portfolios from a broad donor pool, automatically identifying the most influential assets while maintaining interpretability. By employing ℓ_1-regularization rather than explicit cardinality constraints, our approach achieves computational efficiency without sacrificing robustness. The empirical application to S&P 500 constituents shows that relatively few donor assets (27 in our case) are sufficient to create effective synthetic controls.

The copula-based dependence modeling framework proves particularly valuable in capturing complex relationships between target and synthetic assets. Our analysis reveals that elliptical copulas, particularly the Student-t specification, provide the best fit for modeling return dependencies. This finding suggests that while return relationships are predominantly symmetric, they exhibit heavier tails than implied by Gaussian dependence. The superior performance of the Student-t copula over simpler specifications validates the importance of accommodating tail dependence in pairs trading strategies.

The trading strategy's performance metrics across different copula specifications demonstrate the framework's robustness. The N14 mixed copula achieves the highest annualized return (17.26%) and Sharpe ratio (3.97), while maintaining moderate volatility (4.35%). Notably, all specifications deliver positive risk-adjusted returns, with Sharpe ratios ranging from 3.14 to 3.97, suggesting that

the strategy's profitability is not overly dependent on the specific choice of copula.

Several directions for future research emerge from this work. First, exploring time-varying copulas could help capture evolving dependence structures in dynamic market conditions. Second, extending the framework to handle multiple target assets simultaneously could enhance portfolio diversification. Finally, incorporating transaction costs and market impact models would provide more realistic performance estimates for practical implementation.

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.

Table 1: Synthetic Control Model Weights

Tickers	Company Name	Weights (%)
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 Information Services	5.21
PARA	Paramount Global	4.48
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 Corporation	-25.77
ADSK	Autodesk	-42.25
	Total	100.00

Note: This table presents the optimal weights obtained from the sparse synthetic control methodology for replicating the target asset's price dynamics. The weights are expressed as percentages and represent each donor asset's contribution to the synthetic portfolio. Positive weights indicate long positions while negative weights represent short positions. The donor pool consists of S&P 500 constituents, and the methodology yields a sparse solution where many potential donor assets receive zero weights. The sparsity is achieved through ℓ_1-regularization, which automatically selects the most influential assets for constructing the synthetic control. The weights sum to 100% as enforced by the simplex constraint in the optimization problem.

TABLE 2: 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	-1427.05	-1432.94	-1430.81

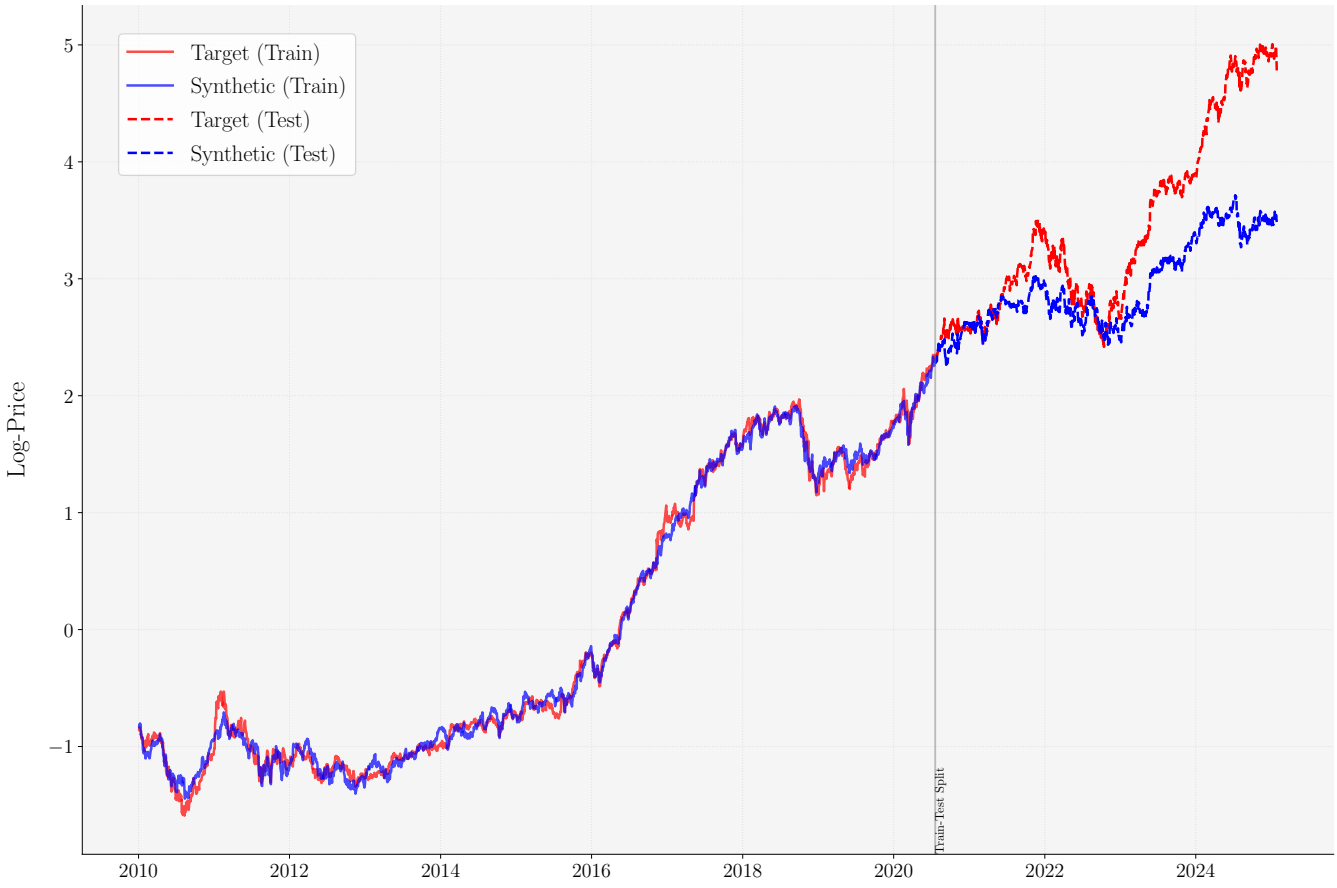
Note: This table reports goodness-of-fit measures for various copula specifications used to model the dependence structure between the target and synthetic asset returns. The evaluation metrics include the Schwarz Information Criterion (SIC), Akaike Information Criterion (AIC), and Hannan-Quinn Information Criterion (HQIC). All criteria balance model fit against complexity, with lower values indicating better models. The Student-t copula achieves the best fit across all three criteria, followed closely by the N14 mixed copula, suggesting that the dependence structure exhibits tail dependence and asymmetry.

TABLE 3: Performance Metrics by Copula

Copula	Total Return (%)	Ann. Return (%)	Ann. Vol. (%)	Sharpe Ratio	Sortino Ratio	Calmar Ratio	Max DD (%)	Win Rate (%)	Profit Factor	VaR-95 (%)
Gumbel	72.59	16.10	4.61	3.49	4.42	7.42	2.17	34.86	2.19	-0.35
Frank	66.53	14.76	3.97	3.71	4.75	10.85	1.36	30.55	2.51	-0.30
Clayton	74.67	16.56	4.18	3.97	5.30	10.89	1.52	32.31	2.60	-0.31
Joe	67.45	14.96	4.62	3.24	3.85	5.83	2.57	36.62	2.02	-0.36
N14	77.82	17.26	4.35	3.97	5.75	11.25	1.53	34.07	2.50	-0.32
Gaussian	62.70	13.91	4.43	3.14	4.10	8.12	1.71	32.31	2.07	-0.35
Student-t	74.63	16.55	4.60	3.60	4.64	7.70	2.15	35.04	2.30	-0.33

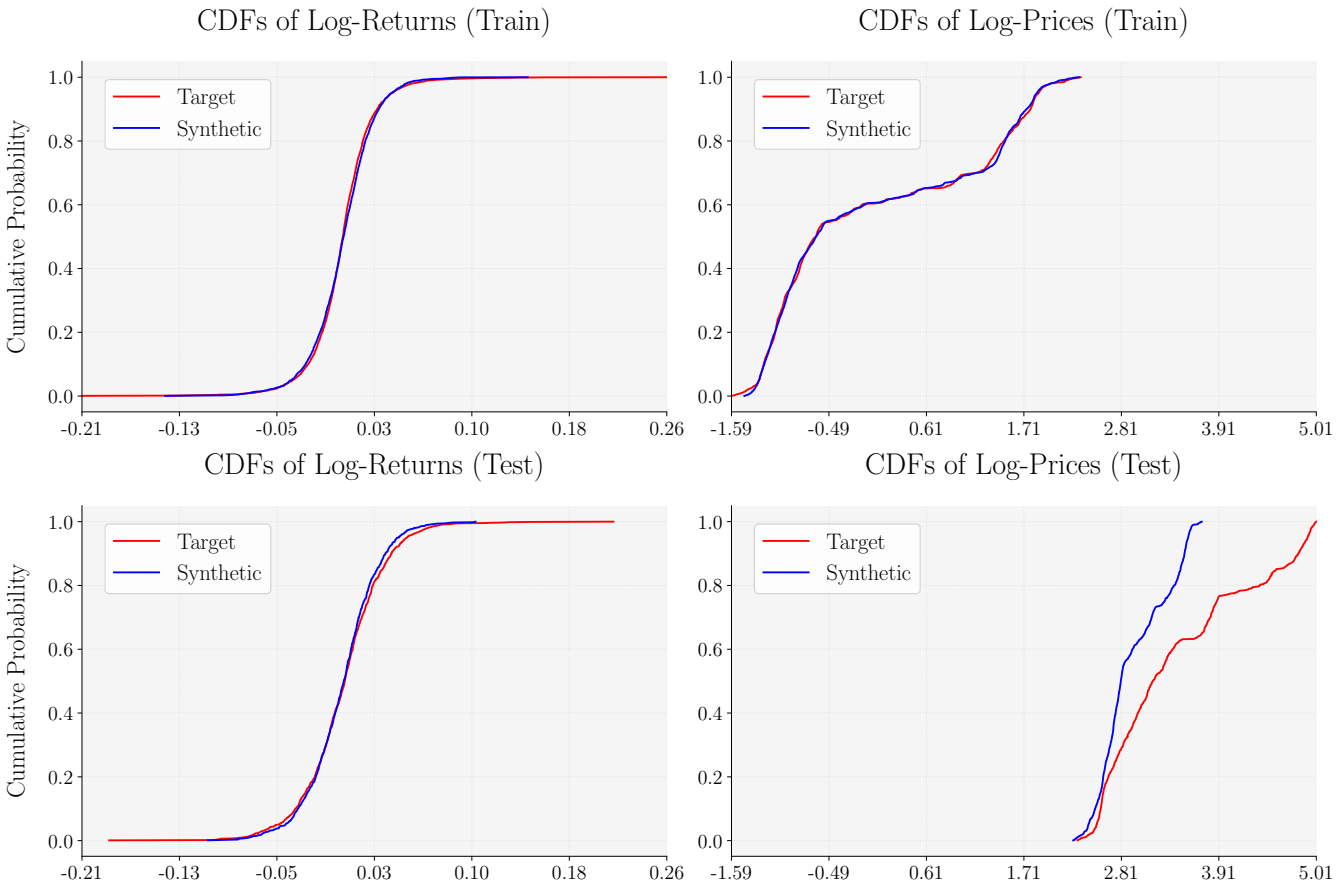
Note: This table reports various performance metrics for pairs trading strategies implemented using different copula specifications. Performance measures include total and annualized returns, annualized volatility, risk-adjusted ratios (Sharpe, Sortino, and Calmar), maximum drawdown (Max DD), win rate, profit factor, and 95% Value-at-Risk (VaR-95). All metrics are computed over the out-of-sample period from July 2020 to January 2025.

FIGURE 1: Target vs Synthetic Log-Prices for NVDA



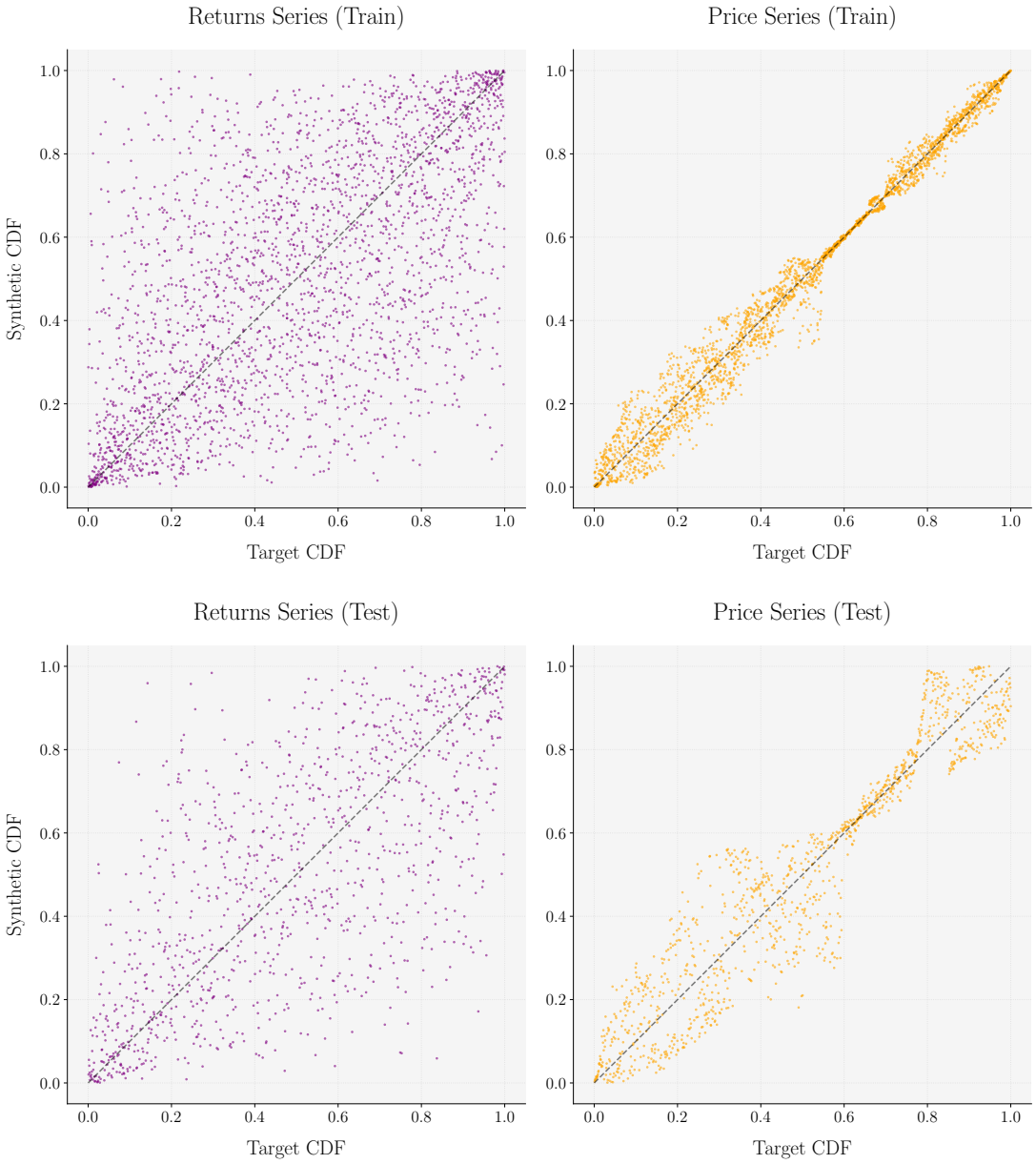
Note: This figure illustrates the log-price trajectories of the target asset (NVDA) and its synthetic counterpart over the training and testing periods. The solid blue line represents the synthetic log-prices derived from the sparse synthetic control methodology, while the solid red line indicates the actual log-prices of the target asset during the training phase. The dashed lines depict the log-prices for both the target and synthetic assets during the testing phase. The vertical line marks the transition point between the training and testing datasets

FIGURE 2: Empirical Cumulative Distribution Functions of Target and Synthetic Assets



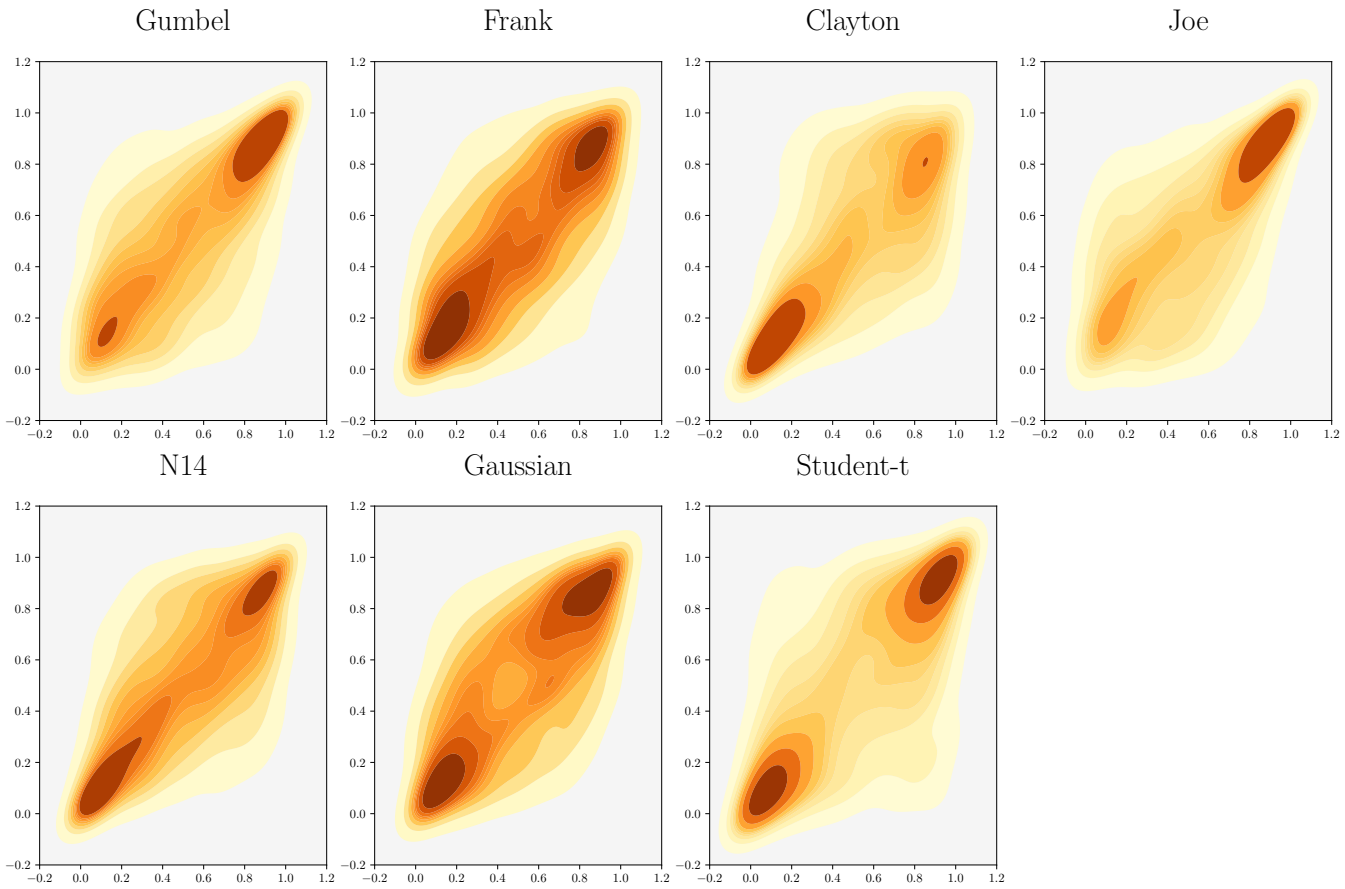
Note: This figure presents the linearly interpolated empirical cumulative distribution functions (ECDFs) for both log-returns and log-prices of the target (red) and synthetic (blue) assets. The panels are divided into training period (top) and test period (bottom). The left panels show the CDFs of daily log-returns, while the right panels display the CDFs of log-price levels. The y-axis represents cumulative probability from 0 to 1, and the x-axis shows the corresponding log-returns or log-prices."

FIGURE 3: Scatter Plots of CDF Pairs for Target and Synthetic Assets



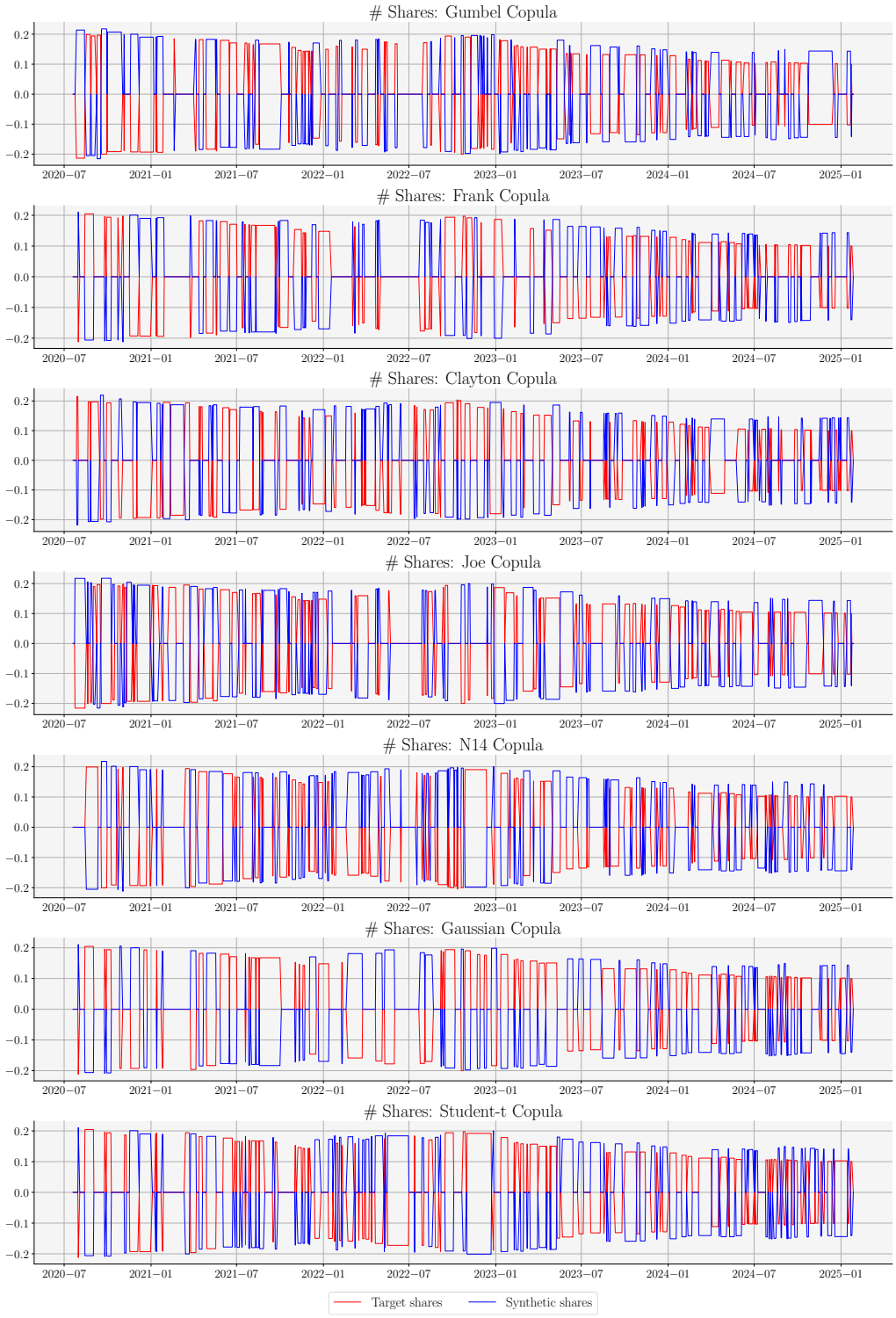
Note: Scatter plots comparing the cumulative distribution functions (CDFs) of the target asset and the synthetic asset for both returns and prices during training and testing periods. The top row displays the CDFs for the training data, while the bottom row shows the CDFs for the testing data. The left column represents the returns series, and the right column represents the price series. Each plot compares the target CDF (x-axis) against the synthetic CDF (y-axis). Both axes represent the [0,1] probability space of the copula domain, with values extending slightly beyond this range to show boundary behavior. The close alignment of points along the diagonal line indicates a strong similarity between the target and synthetic distributions. The returns series exhibit more dispersion, reflecting the volatility in returns, while the price series show a tighter fit, suggesting that the synthetic asset effectively replicates the price dynamics of the target asset.

FIGURE 4: Copula Density Heatmaps for Different Parametric Families



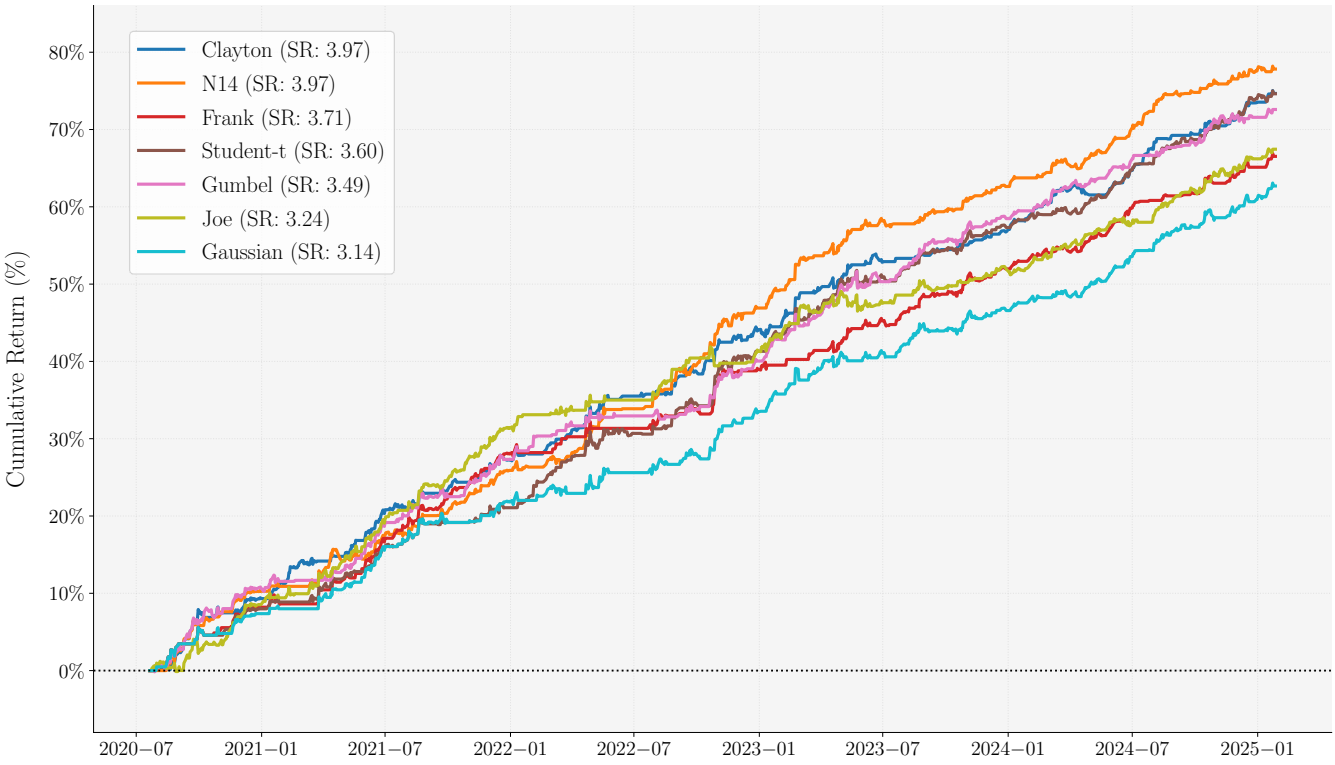
Note: Contour plots of copula densities for various parametric families, illustrating the dependence structures between the target and synthetic asset returns. Each subplot represents a different copula family: Gumbel, Frank, Clayton, Joe, N14, Gaussian, and Student-t. The color intensity in the plots indicates the density of the copula, with darker regions representing higher densities. These visualizations highlight the differences in tail dependencies and overall dependence structures captured by each copula family. The Gumbel and Clayton copulas exhibit stronger upper and lower tail dependencies, respectively, while the Gaussian copula shows symmetric dependencies. The Student-t copula accounts for heavier tails, and the Frank copula displays a balanced dependence structure.

FIGURE 5: Position Sizes Over Time by Copula Family



Note: This figure shows the evolution of position sizes (in shares) for target (red) and synthetic (blue) assets in the test sample under different copula specifications. The strategy assumes initial equity of $1.

FIGURE 6: Equity Curves Comparison Across Copula Families



Note: This figure compares the cumulative returns of trading strategies based on different copula specifications in the test sample. Each line represents a different copula model, with their respective Sharpe Ratios (SR) shown in parentheses. The y-axis shows cumulative returns in percentage terms, and the x-axis displays the timeline in six-month intervals.

A. Appendix

A.1 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 $i \in \{1, \ldots, N\}$. 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* $\{y_1, y_2, \ldots, y_n\}$ *be a collection of random variables, where* y_1 *is the "target" variable and* $\mathbf{y}_{2:n} = (y_2, \ldots, y_n)$ *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 $\kappa = (\beta_1)^{-1} = -(\sum_{j=2}^{n} \beta_j)^{-1}$

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

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

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

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. □

A.2 Why not use a cardinality-constrained Synthetic Control?

While the ℓ_1-regularized approach provides a computationally efficient and convex framework for constructing sparse synthetic controls, it is worth considering alternative methods that directly impose sparsity through cardinality constraints. A natural alternative is to solve a cardinality-constrained quadratic program, which explicitly limits the number of non-zero weights in the synthetic asset. Formally, this can be expressed as:

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

where $\|\mathbf{w}\|_0 := \sum_{i=1}^N \mathbb{I}\{w_i \neq 0\}$ counts the number of non-zero elements in $\mathbf{w}$, and K is a user-defined sparsity level. This formulation directly enforces sparsity by restricting the synthetic asset to be constructed from at most K donor assets. However, the cardinality constraint introduces significant computational challenges, as the problem becomes NP-hard due to its combinatorial nature. Below, we discuss two approaches to approximate this problem and their limitations.

A.2.1 Mixed-Integer Programming Approach

One way to tackle the cardinality-constrained problem is to reformulate it as a mixed-integer quadratic program (MIQP). This involves introducing binary variables $z_i \in \{0, 1\}$ for $i = 1, \ldots, N$, where $z_i = 1$ indicates that the i-th asset is included in the synthetic control, and $z_i = 0$ otherwise. The problem can then be rewritten as:

$$\mathbf{w}^*, \mathbf{z}^* = \left[\begin{array}{ll} \underset{\mathbf{w} \in \mathbb{R}^N,\ \mathbf{z} \in \{0,1\}^N}{\arg\min} & \sum_{t=1}^T \left(y_t - \sum_{i=1}^N w_i x_{it} \right)^2 \\ \text{s.t.} & \left| \begin{array}{ll} \mathbf{1}^\top \mathbf{w} & = 1, \\ \sum_{i=1}^N z_i & \leq K, \\ |w_i| & \leq M z_i \quad \text{for } i = 1, \ldots, N, \end{array} \right. \end{array} \right]$$

where M is a sufficiently large constant that bounds the magnitude of the weights. The constraint $|w_i| \leq M z_i$ ensures that w_i can only be non-zero if $z_i = 1$. While this formulation is exact, it is computationally intensive, especially for large donor pools, as it requires solving a mixed-integer program. The computational complexity grows exponentially with the number of assets, making it impractical for high-dimensional settings.

A.2.2 Two-Step Heuristic Procedure

An alternative approach is to use a two-step heuristic procedure that approximates the cardinality-constrained solution without requiring mixed-integer programming. This procedure proceeds as follows:

1. **Solve the full least squares problem:** First, solve the unconstrained least squares problem to obtain an initial weight vector:

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

2. **Select the K largest weights:** Identify the K largest weights (in absolute value) from $\mathbf{w}^{(1)}$ and define the support set:

$$\mathcal{I} := \{i : |w_i^{(1)}| \text{ is among the } K \text{ largest}\}.$$

3. **Solve the restricted program:** Solve the least squares problem restricted to the support set $\mathcal{I}$:

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

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

4. **Construct the full weight vector:** Embed the optimized restricted weights back into the original N-dimensional space:

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

While this heuristic is computationally efficient, it has several drawbacks. First, the initial least squares solution $\mathbf{w}^{(1)}$ may not provide a good indication of which assets are most relevant, especially in the presence of multicollinearity or noise. Second, the procedure can lead to extreme

weights (both positive and negative) in the final solution, resulting in a highly leveraged portfolio that may not be practical for trading. This is because the restricted optimization step does not impose any bounds on the magnitude of the weights, allowing for large positive and negative values that cancel each other out to satisfy the unit sum constraint.

A.2.3 Comparison with ℓ_1-Regularized Approach

In contrast to the cardinality-constrained approaches, the ℓ_1-regularized method provides a more balanced trade-off between sparsity and computational efficiency. By shrinking some weights exactly to zero, the ℓ_1 penalty achieves sparsity without requiring explicit cardinality constraints. Moreover, the convex nature of the problem ensures that it can be solved efficiently using proximal algorithms or quadratic programming techniques, even for high-dimensional donor pools. Additionally, the regularization parameter λ provides fine-grained control over the sparsity level, allowing the user to tune the solution based on their specific requirements.

In practice, we found that the ℓ_1-regularized approach yields more stable and interpretable synthetic controls compared to the cardinality-constrained methods. The latter often produce highly leveraged portfolios with extreme weights, which are undesirable in a trading context. Furthermore, the computational advantages of the ℓ_1-regularized approach make it more suitable for real-world applications, where scalability and robustness are critical.

In conclusion, while cardinality-constrained formulations offer a conceptually appealing way to enforce sparsity, their practical limitations make them less attractive for constructing synthetic controls in pairs trading. The ℓ_1-regularized approach strikes a better balance between sparsity, interpretability, and computational efficiency, making it the preferred choice for our application.

A.3 Algorithms

Algorithm 1. L1-Regularized Synthetic Control with TimeSeriesSplit Cross-Validation

Require:

Target asset log-prices $\mathbf{y} = [y_t]_{t=1}^{T} \in \mathbb{R}^T$

Donor pool log-prices $\mathbf{X} = [x_{1t}, ..., x_{Nt}]_{t=1}^{T} \in \mathbb{R}^{T \times N}$

Candidate regularization parameters $\Lambda = \{\lambda_1, ..., \lambda_M\}$

Number of time series splits n_splits

Ensure:

Optimal sparse weight vector $\mathbf{w}^* \in \mathbb{R}^N$

function TimeSeriesSplit(T, n_splits) ▷ Extract expanding-window train/validation sets

  $n \leftarrow \lfloor T/(n\_splits + 1) \rfloor$ ▷ Length of test fold

  for $i = 1$ to n_splits **do**

    $train\_end \leftarrow n \cdot (i + 1)$ ▷ End index of training set

    $test\_end \leftarrow n \cdot (i + 2)$ ▷ End index of validation set

    return $[1 : train\_end], [train\_end + 1 : test\_end]$ ▷ Train/test indices

  end for

end function

function SyntheticControl($\mathbf{y}, \mathbf{X}, \Lambda, n\_splits$)

  for each $\lambda \in \Lambda$ **do** ▷ for each λ, perform time series cross-validation

    $mse_\lambda \leftarrow 0$

    for $(train\_idx, val\_idx)$ in TimeSeriesSplit(T, n_splits) **do**

      $\mathbf{X}_{train} \leftarrow \mathbf{X}[train\_idx], \mathbf{y}_{train} \leftarrow \mathbf{y}[train\_idx]$ ▷ Extract training data

      $\mathbf{X}_{val} \leftarrow \mathbf{X}[val\_idx], \mathbf{y}_{val} \leftarrow \mathbf{y}[val\_idx]$ ▷ Extract validation data

      $\mathbf{w}_\lambda = \arg\min_{\mathbf{w}} \{\|\mathbf{y}_{train} - \mathbf{X}_{train}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_1\}$ s.t. $\mathbf{1}^\top\mathbf{w} = 1$ ▷ Solve ℓ_1-program

      $mse_\lambda \leftarrow mse_\lambda - \|\mathbf{y}_{val} - \mathbf{X}_{val}\mathbf{w}_\lambda\|_2^2$ ▷ Accumulate negative MSE for scoring

    end for

    $score_\lambda \leftarrow mse_\lambda / n\_splits$ ▷ Average neg. MSE

  end for

  $\lambda^* \leftarrow \arg\max_{\lambda \in \Lambda} \{score_\lambda\}$

  $\mathbf{w}^* = \arg\min_{\mathbf{w}} \{\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda^*\|\mathbf{w}\|_1\}$ s.t. $\mathbf{1}^\top\mathbf{w} = 1$

  return $\mathbf{w}^*, \mathcal{I}^*$

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: $x_{(1)} < \cdots < x_{(m)}$

        $p_i \leftarrow \frac{1}{T-1} \sum_{t=2}^{T} \mathbb{I}(x_t \le x_{(i)})$ ▷ 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 ϵ

        $v_t \leftarrow \max\{\epsilon, \min\{\hat{F}_{R^*}(r_t^*), 1 - \epsilon\}\}$

    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)

  $MI_t^{R|R^*} \leftarrow \dfrac{\partial C_{\hat{\theta}}(u_t, v_t)}{\partial v_t}$ ▷ Compute target-synthetic MI

  $MI_t^{R^*|R} \leftarrow \dfrac{\partial C_{\hat{\theta}}(u_t, v_t)}{\partial u_t}$ ▷ 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.4 Barra model

The Barra model for our target and synthetic asset may be written as

$$\begin{bmatrix} r_t \\ r_t^* \end{bmatrix} = \begin{bmatrix} \alpha \\ \alpha^* \end{bmatrix} + \begin{bmatrix} \boldsymbol{\beta}^\top \\ \boldsymbol{\beta}^{*\top} \end{bmatrix} \mathbf{f}_t + \begin{bmatrix} \boldsymbol{\gamma}^\top \\ \boldsymbol{\gamma}^{*\top} \end{bmatrix} \mathbf{i}_t + \begin{bmatrix} \epsilon_t \\ \epsilon_t^* \end{bmatrix}$$

where we consider $K = 8$ fundamental factors $\mathbf{f}_t$ (i.e.: $\boldsymbol{\beta}, \boldsymbol{\beta}^*, \mathbf{f}_t \in \mathbb{R}^K$) and $M = 17$ industry factors $\mathbf{i}_t$ (i.e.: $\boldsymbol{\gamma}, \boldsymbol{\gamma}^*, \mathbf{i}_t \in \mathbb{R}^M$). The "*active return*" between the target and synthetic asset is given by:

$$\dot{r}_t := r_t - r_t^* = (\alpha - \alpha^*) + (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top \mathbf{f}_t + (\boldsymbol{\gamma} - \boldsymbol{\gamma}^*)^\top \mathbf{i}_t + (\epsilon_t - \epsilon_t^*).$$

Now defining the *relative alpha, beta and gamma*, respectively, as $\dot{\alpha} := (\alpha - \alpha^*), \dot{\boldsymbol{\beta}} := (\boldsymbol{\beta} - \boldsymbol{\beta}^*), \dot{\boldsymbol{\gamma}} := (\boldsymbol{\gamma} - \boldsymbol{\gamma}^*)$ and setting $\dot{\epsilon}_t := (\epsilon_t - \epsilon_t^*)$, we may write the model in terms of the portfolio's active return

$$\dot{r}_t = \dot{\alpha} + \dot{\boldsymbol{\beta}}^\top \mathbf{f}_t + \dot{\boldsymbol{\gamma}}^\top \mathbf{i}_t + \dot{\epsilon}_t. \tag{3}$$

In Figure A7 we show the factor correlation matrix $\text{Corr}(\mathbf{X}) \in \mathbb{R}^{J \times J}$ of all the factors

$$\mathbf{X} = \begin{bmatrix} \mathbf{f}_1^\top, \mathbf{i}_1^\top \\ \vdots \quad \vdots \\ \mathbf{f}_T^\top, \mathbf{i}_T^\top \end{bmatrix} \in \mathbb{R}^{T \times J},$$

where $J = K + M$. In our application we are using [`MKT_RF`, `SMB`, `HML`, `RMW`, `CMA`, `MOM`, `ST_REV`, `LT_REV`] as the fundamental factors, and [`Food`, `Mines`, `Oil`, `Clths`, `Durbl`, `Chems`, `Cnsum`, `Cnstr`, `Steel`, `FabPr`, `Machn`, `Cars`, `Trans`, `Utils`, `Rtail`, `Finan`, `Other`] as the industry factors. As we can see, correlations are very high among factors, which means that regular OLS estimation of eq. (3) will deliver highly unstable coefficients due to multicollinearity.

[Insert Figure A7 about here]

Hence, to properly estimate the model parameters, we employ an orthogonal regression approach based on Principal Component Analysis (PCA), which will allow us to obtain more stable estimates of the factor exposures. The implementation follows these steps.

First, we compute the covariance matrix of the factors $\boldsymbol{\Sigma} := \text{Cov}(\mathbf{X}) \in \mathbb{R}^{J \times J}$ and obtain its eigendecomposition $\boldsymbol{\Sigma}\mathbf{V} = \boldsymbol{\Lambda}\mathbf{V}$, where $\boldsymbol{\Lambda} := \text{diag}(\lambda_1, ..., \lambda_J) \in \mathbb{R}^{J \times J}$ are the eigenvalues and $\mathbf{V} := [\mathbf{v}_1, ..., \mathbf{v}_J] \in \mathbb{R}^{J \times J}$ are the corresponding eigenvectors, both sorted in descending order of the λ's. The principal components are given by: $\mathbf{P} = \mathbf{X}\mathbf{V} \in \mathbb{R}^{T \times J}$.

Second, we regress the active returns onto the principal components

$$\dot{r}_t = a^{(u)} + \sum_{i=1}^{J} b_i^{(u)} p_{t,i} + \nu_t^{(u)}$$

where $a^{(u)}$ is the "*unrestricted*" intercept, $\boldsymbol{b}^{(u)} := [b_1^{(u)}, ..., b_J^{(u)}]$ are the "*unrestricted*" coefficients for each principal component, and ν_t is the error term. We keep only the statistically significant principal components at the 0.05 significance level: $\mathcal{S} := \{i : p\text{-value}(b_i^{(u)}) < 0.05\}$. Then, we estimate a restricted model using only the significant principal components

$$\dot{r}_t = a^{(r)} + \sum_{i \in \mathcal{S}} b_i^{(r)} p_{t,i} + \nu_t^{(r)}.$$

Finally, we transform the coefficients back to original factor space. Let $\boldsymbol{b}^{(r)} \in \mathbb{R}^J$ denote the vector filled with $b_i^{(r)}$ if $i \in \mathcal{S}$ and 0 otherwise. Then, we can write $\dot{r}_t = a^{(r)} + \mathbf{p}_t^\top \boldsymbol{b}^{(r)} + \nu_t^{(r)} = a^{(r)} + \mathbf{x}_t^\top \mathbf{V} \boldsymbol{b}^{(r)} + \nu_t^{(r)}$, where $\mathbf{p}_t$ and $\mathbf{x}_t$ are rows of $\mathbf{P}$ and $\mathbf{X}$, respectively (given as column vectors). Thus, by setting $\dot{\alpha} = a^{(r)}$ and $[\dot{\boldsymbol{\beta}}\ \dot{\boldsymbol{\gamma}}]^\top = \mathbf{V}\boldsymbol{b}^{(r)}$ we recover alpha and the factor betas and gammas while avoiding the instability due to multicollinearity. Both the unrestricted and restricted models are estimated with Heteroskedasticity and Autocorrelation Consistent (HAC) standard errors using a maximum lag of 5 periods to account for potential serial correlation and heteroskedasticity in the residuals.

This approach offers several advantages. First, by using orthogonal principal components, we eliminate multicollinearity concerns. Second, by selecting only significant components, we reduce dimensionality and potential overfitting. Finally, the transformation back to the original factor space allows for direct interpretation of the factor exposures $\dot{\boldsymbol{\beta}}$ and $\dot{\boldsymbol{\gamma}}$ in our active return decomposition model.

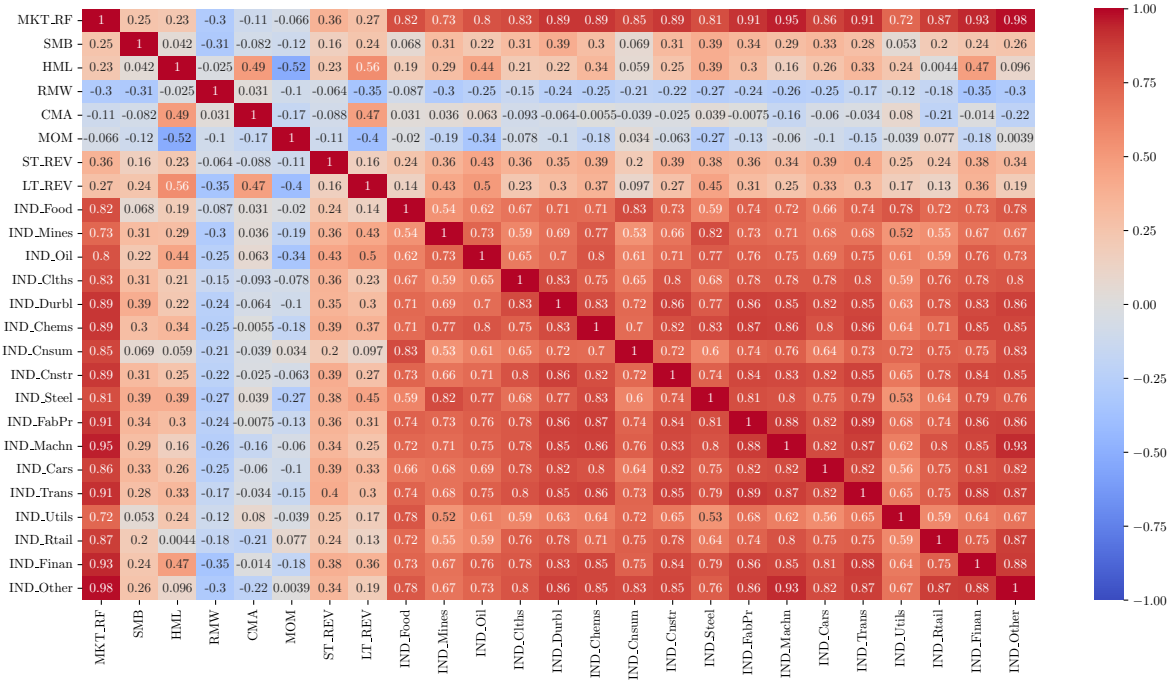
TABLE A4: Factor Model Regression Results for Pairs Trading Strategy

	train	**test**
Alpha and Significance		
Alpha	-0.0002	0.0009
Alpha SE	(0.0004)	(0.0006)
Alpha p-value	[0.6766]	[0.1419]
Factor Betas		
β_{MKT_RF}	0.0426	0.0281
β_{SMB}	0.3802	0.0285
β_{HML}	-0.1773	-0.4563
β_{RMW}	0.6985	0.3470
β_{CMA}	0.3291	0.1318
β_{MOM}	-0.0215	-0.2389
β_{ST_REV}	0.1232	0.0714
β_{LT_REV}	0.1855	0.4929
β_{Food}	0.0312	0.0459
β_{Mines}	-0.0431	0.0071
β_{Oil}	-0.0291	-0.1447
β_{Clths}	0.1369	0.0364
β_{Durbl}	-0.2761	-0.1056
β_{Chems}	-0.1755	-0.1059
β_{Cnsum}	-0.0545	-0.0683
β_{Cnstr}	0.2890	-0.0143
β_{Steel}	-0.3551	-0.4512
β_{FabPr}	-0.1014	-0.1088
β_{Machn}	0.4337	1.0429
β_{Cars}	0.0899	0.0027
β_{Trans}	-0.1163	0.2022
β_{Utils}	0.1995	0.2607
β_{Rtail}	-0.5772	-0.6114
β_{Finan}	0.3596	0.2546
β_{Other}	0.0215	-0.1835
Model Statistics		
Adj. R^2	0.0687	0.2525
$F - statistic$	10.5893	19.7223
$Fp - value$	0.0000	0.0000

Notes: This table reports the regression results for the Pairs Trading Strategy using a PCA-based factor model. The alpha represents the abnormal return after controlling for factor exposures. Standard errors are in parentheses and p-values are in brackets. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

FIGURE A7: Factor Correlation Matrix

(A) Train



(B) Test