Rapport Pair Trading

Une stratégie quantitative sur le S&P500

Lancelot Watelle

Janvier 2026

Table des matières

1 Introduction

1.1 Qu'est-ce que le pair trading ?

Le pair trading, bien qu'étant une stratégie quantitatives ancienne, reste aujourd'hui largement utilisé dans le secteur financier.

On peut le définir comme une **stratégie neutre** qui permet de tirer profit de la plupart des mouvements de marché (hausse, baisse, etc.). Cette approche consiste à surveiller la performance de deux titres historiquement corrélés [1]. Lorsque cette corrélation s'affaiblit temporairement — c'est-à-dire qu'un titre monte tandis que l'autre baisse — la stratégie consiste à **vendre à découvert le titre le plus performant** et à **acheter le titre le moins performant**, en pariant que cet écart finira par converger.

Étant une stratégie historiquement utilisée et à la base de l'arbitrage statistique, j'ai souhaité tester une implémentation quantitative sur l'indice mondialement reconnu : le S&P500.

1.2 Pourquoi le S&P500 ?

La bourse américaine est le marché le plus volumineux et dynamique du monde. Elle représente plus de 50 % de la capitalisation boursière mondiale et près de 40 % des transactions financières globales.

Son indice principal, le S&P500, regroupe les 500 plus grandes entreprises cotées aux États-Unis. Avec sa grande diversité de secteurs et d'industries, il permet d'identifier des entreprises présentant des relations statistiques fortes tout en les regroupant par secteur.

Le S&P500 constitue donc un indice idéal pour repérer des paires d'actions susceptibles de présenter une cointégration significative.

1.3 Objectif du projet

L'objectif de ce projet est d'étudier la cointégration des actifs du S&P500 et de construire un portefeuille fictif permettant la mise en place d'une stratégie de pair trading optimisée.

1. La cointégration est une propriété statistique des séries temporelles permettant de détecter une relation de long terme entre deux ou plusieurs séries.

2 Méthodologie

2.1 Données

Durant l'ensemble de ce projet, nous utilisons les données de clôture du marché quotidienne (jours ouvrés). L'analyse qu'exige une telle stratégie justifie ce choix.

L'ensemble des données provient de l'API gratuite de Python *yfinance*. Sa simplicité et sa fiabilité ont motivé ce choix.

La période analysée dépend de l'étape du projet. Pour les tests globaux, nous avons retenu une durée de 2-5 ans. Cependant, lors de la validation de notre stratégie, nous faisons varier cette période pour analyser sa robustesse.

Pour récupérer les informations relatives aux entreprises du S&P500, nous utilisons le web scraping (respectant l'ensemble des réglementations) sur le site de Wikipedia, ce qui nous permet de récupérer les entreprises, leurs symboles, leurs secteurs d'activité et leurs sous-catégories.

2.2 Qu'est-ce que le pair trading mathématiquement ?

Le pair trading vise à parier sur l'équilibre des écarts entre deux actifs. Si nous décrivons ce phénomène mathématiquement, l'équation fondamentale liant les deux actifs est :

$$\text{actif}_1 = \beta \cdot \text{actif}_2 + \text{spread} \tag{1}$$

Nous analysons la stationnarité [2] du *spread* en attendant une perturbation pour prendre une position. L'idée centrale est que lorsque le *spread* s'écarte de sa valeur d'équilibre, nous pouvons construire une position profitable en anticipant son retour à la moyenne.

2. Une série stationnaire est une série temporelle dont les propriétés statistiques (moyenne, variance) ne dépendent pas du temps

3 Construction de la stratégie

3.1 Walk-Forward Analysis

Pour évaluer l'efficacité d'une paire sur une période donnée, nous avons adopté l'approche **Walk-Forward Analysis**. Cette méthode divise une période en deux phases distinctes :

— **Phase 1 (historique)** : étude des données pour calculer les indicateurs et les paramètres

— **Phase 2 (forward)** : trading réel en fonction des paramètres préalablement calculés

Cette approche nous permet d'optimiser les performances en choisissant une période de trading courte tout en analysant un grand volume de données historiques. Nous avons donc choisi une **période d'analyse de 1 an** pour une **période de trading de 1 mois**.

3.2 Paramètres d'analyse

3.2.1 Différence de poids (Risk-Adjusted Weighting)

Dans la méthode classique du pair trading, la pondération dépend uniquement du coefficient bêta de régression. Cependant, pour diminuer le risque tout en améliorant les performances, nous utilisons la volatilité passée d'un actif pour pondérer notre équation.

L'équation devient :

$$\frac{1}{\text{vol}_{\text{actif}_1}} \cdot \text{actif}_1 = \frac{\beta}{\text{vol}_{\text{actif}_2}} \cdot \text{actif}_2 + \text{spread} \tag{2}$$

où la volatilité représente l'écart-type des rendements quotidiens en valeur absolue.

Les pondérations finales sont calculées comme :

$$\text{pondération}_{\text{actif}_1} = \frac{\frac{1}{\text{vol}_{\text{actif}_1}}}{\frac{1}{\text{vol}_{\text{actif}_1}} + \frac{\beta}{\text{vol}_{\text{actif}_2}}} \tag{3}$$

$$\text{pondération}_{\text{actif}_2} = 1 - \text{pondération}_{\text{actif}_1} \tag{4}$$

Cette approche permet une allocation plus équilibrée en tenant compte du profil de risque de chaque actif.

3.2.2 Volume et Liquidité

Le volume d'un actif est un indicateur crucial de sa liquidité. Dans une stratégie de pair trading, la liquidité est essentielle pour pouvoir entrer et sortir de positions sans impact de marché significatif.

Nous définissons l'**Average Daily Dollar Volume** (ADDV) comme :

$$\text{ADDV} = \frac{\sum_{i=1}^{n}(\text{Prix}_i \times \text{Volume}_i)}{n} \tag{5}$$

Ce paramètre nous permet de filtrer les paires ayant une liquidité insuffisante. Chaque mois, nous ne conservons que les 20 % des paires ordonnées par ADDV décroissant.

3.2.3 Test de cointégration

Une série temporelle est une suite de valeurs numériques représentant l'évolution d'une quantité au cours du temps. La modélisation des cours financiers repose largement sur ce domaine, et un concept clé est la **cointégration**.

La cointégration est une propriété statistique qui détecte les relations de long terme entre plusieurs séries temporelles. Il est important de distinguer :

— **Corrélation** : détecte les similitudes sur les variations (si A augmente, B augmente aussi)

— **Cointégration** : étudie si la différence entre les séries reste stable à long terme

Pour évaluer si deux paires sont cointégrées, nous utilisons le test de **Dickey-Fuller augmenté** sur le résidu de régression.

Définition formelle : Deux séries y_t et x_t, intégrées d'ordre 1 $[I(1)]$, sont cointégrées s'il existe un vecteur $\beta = (1, -\beta_1)$ tel que :

$$u_t = y_t - \beta_1 x_t \sim I(0)$$

où u_t est stationnaire. Étudier la cointégration de deux paires est donc essentiel pour une stratégie de pair trading efficace.

3.2.4 Demi-vie du spread (Mean Reversion Speed)

La *demi-vie de retour à la moyenne* représente le temps nécessaire pour qu'un spread perturbé s'ajuste de moitié vers sa valeur d'équilibre. Ce paramètre est crucial pour dimensionner nos positions et gérer le risque.

Nous utilisons le processus stochastique d'**Ornstein-Uhlenbeck** :

$$ds_t = \lambda(\mu - s_t)\, dt + \sigma\, dW_t \tag{6}$$

où :

— s_t : spread

— λ : vitesse de retour vers la moyenne

— μ : moyenne à long terme

La solution de ce processus montre que :

$$\text{demiVie} = \frac{\ln(2)}{\lambda} \tag{7}$$

En pratique, nous estimons λ par régression linéaire discrète :

$$\Delta s_t = a \cdot s_{t-1} + b + \epsilon_t \quad \text{où} \quad a = -\lambda \tag{8}$$

3.2.5 Volatilité du spread

Maintenant que nous savons si le spread reviendra à sa moyenne rapidement ou non, il est important d'évaluer son amplitude de variation. Nous utilisons simplement l'**écart-type du spread** sur la période d'analyse de données.

3.2.6 Pseudo-trading (Backtesting on Historical Data)

Pour prédire la performance d'une paire en trading réel, nous simulons un trading fictif sur les 6 derniers mois de notre période d'analyse. Cette période est divisée en 6 sous-périodes de 1 mois pour être aussi réaliste que possible.

Les paramètres extraits sont :

— **Average PnL** : moyenne des PnL samples
— **Prop PnL Pos** : proportion des périodes gagnantes (%)
— **Worst PnL** : pire résultat mensuel
— **Avg Max Drawdown** : moyenne des drawdowns mensuels

4 Analyse des données

4.1 Approche initiale vs Approche retenue

Initialement, j'avais envisagé de créer une régression linéaire entre mes paramètres et le PnL pour prédire les performances futures. Cependant, je me suis rapidement rendu compte que certaines paires extraordinairement performantes faussaient complètement le modèle, rendant le R^2 proche de 0 (régression inutile).

J'ai donc adopté une approche alternative basée sur le **ranking** : classer les paires selon l'influence relative de leurs paramètres sur le trading du mois suivant.

4.2 Interprétation par quantile

L'**analyse par segmentation en quantiles** est une méthode statistique permettant de comprendre les relations non-linéaires entre variables.

La méthode se divise en 3 étapes :

1. **Ordonnancement** : on ordonne les données selon la variable à expliquer (ici le PnL du mois)
2. **Division** : on sépare l'échantillon en quantiles (nous avons choisi 5 quantiles)
3. **Analyse** : pour chaque variable explicative, on calcule sa valeur moyenne par quantile et on cherche une tendance

4.3 Création des données

Sur la période **2018-01-01 à 2022-01-01** (48 périodes), nous calculons les indicateurs pour chaque paire d'un même secteur et le PnL réel du mois suivant. Cela crée un dataset permettant d'analyser les relations entre indicateurs et performance.

4.4 Z-scores et normalisation

Pour comparer des variables provenant de distributions différentes, nous normalisons chaque variable en score standardisé :

$$\text{z\_score}_{\text{element}} = \frac{\text{variable}_{\text{element}} - E[\text{variable}]}{\sigma[\text{variable}]} \tag{9}$$

Dans certains cas, nous plafonnons les valeurs extrêmes (cap max à 1) pour éviter qu'une variable unique ne domine le ranking. Pour les variables avec une répartition en U-shape (où les valeurs centrées sont meilleures), nous utilisons :

$$z_{\text{final}} = \exp\left(-\frac{z^2_{\text{variable}}}{2}\right) \tag{10}$$

4.5 Information Coefficient (IC Score)

L'IC mesure la qualité prédictive d'une variable pour le PnL du mois suivant :

$$\text{IC} = \text{corr}(\text{zscore}_{\text{var}}, \text{pnl\_month\_suivant}) \tag{11}$$

En calculant l'IC pour chaque variable, nous obtenons une mesure de son importance relative.

4.6 Fonction finale de ranking

Bien que certaines variables (p-value, volume) aient un IC faible, elles restent utiles comme **filtres** :

— **Filtre 1** : ne conserver que les paires avec p-value < 0.05 (paires cointégrées)

— **Filtre 2** : ne conserver que les 20 % des paires avec le plus grand ADDV

Pour les variables restantes, nous créons un ranking pondéré :

$$\alpha_i = \frac{|\mathrm{IC}_i|}{\sum |\mathrm{IC}|} \tag{12}$$

$$\mathrm{RANKING(paire)} = \sum_{i=1}^{n} \alpha_i \cdot z\_\mathrm{score}_{\mathrm{paire,i}} \tag{13}$$

Pour éviter les biais de performance passée, nous pénalisons les paires ayant performé au mois précédent :

$$\mathrm{RANKING\_FINAL(paire)} = \frac{\mathrm{RANKING(paire)}}{1 + 0.5 \cdot \mathrm{count}_{m-1} + 0.30 \cdot \mathrm{count}_{m-2}} \tag{14}$$

5 Backtest de la stratégie

5.1 Architecture de la stratégie

La stratégie se sépare en deux étapes mensuelles :

Étape 1 : Analyse des données (fin du mois précédent)

1. Récupérer pour chaque paire du même secteur ses indicateurs (basés sur les 12 derniers mois)
2. Appliquer les filtres (p-value, ADDV)
3. Ordonner les lignes avec la fonction de ranking
4. Extraire les 20 meilleures paires en respectant : max 2 fois le même stock, max 40 % par secteur

Étape 2 : Trading (durant le mois)

1. Traiter chaque paire selon la stratégie de pair trading (avec les paramètres calculés)
2. Allouer $\frac{\text{value\_ptf}}{nb\_paires}$ à chaque paire
3. À la fin du mois, forcer la sortie de toutes les positions
4. Réinvestir le portefeuille (avec ou sans intérêts composés)

6 Performances réelles

6.1 Backtesting sur 2022-2026

Le modèle a été calibré sur 2018-2022, puis testé sur 2022-2025. Les résultats montrent des performances encourageantes avec deux approches de gestion des intérêts :

6.1.1 Avec composition des intérêts

Le portefeuille affiche une croissance constante, avec un Sharpe ratio moyen de **1.35** sur la période. Les performances annuelles sont :

Table 1 – Performances annuelles avec composition des intérêts

Année	Final Value	Profitability	Max DD	Volatility	Sharpe
2022-2023	224 330	49.55 %	-4.54 %	4.64 %	2.40
2023-2024	168 630	12.42 %	-4.39 %	2.94 %	0.85
2024-2025	163 966	9.31 %	-1.37 %	2.05 %	0.79
Global	185 642	23.76 %	-3.43 %	3.21 %	1.35

La période 2022-2023 a particulièrement bien performé, impactant positivement l'ensemble des indicateurs. Le max drawdown reste très faible (moyenne -3.5 %), ce qui montre la robustesse de la stratégie.

Performance mensuelle :

— 75 % de mois positifs
— 25 % de mois négatifs
— Moyenne mensuelle : 1.76 %

6.1.2 Sans composition des intérêts

L'approche sans réinvestissement des gains montre un profil risque-rendement légèrement différent :

Table 2 – Performances annuelles sans composition des intérêts

Année	Final Value	Profitability	Max DD	Volatility	Sharpe
2022-2023	214 340	42.89 %	-2.52 %	3.80 %	2.54
2023-2024	169 239	12.83 %	-3.03 %	2.59 %	1.00
2024-2025	164 500	9.67 %	-1.25 %	1.95 %	0.88
Global	182 693	21.80 %	-2.27 %	2.78 %	1.47

Le risque est logiquement moins important (max drawdown baisse à -2.3 %), tandis que le PnL total diminue légèrement. Le Sharpe ratio s'améliore à 1.47.

6.2 Beta de la stratégie

Un indicateur clé est le **bêta** de la stratégie, mesurant sa sensibilité aux variations du marché global.

$$\beta = \frac{\mathrm{Corr}(R_{\text{stratégie}}, R_{\text{S\&P500}})}{\mathrm{Var}(R_{\text{S\&P500}})} = 0.0049 \tag{15}$$

Ce bêta quasi-nul confirme que notre stratégie est **market-neutral**, ce qui est logique pour une stratégie purement quantitative basée sur la cointégration statistique.

7 Problèmes rencontrés et solutions

7.1 Prédiction vs Ranking

Problème : La régression linéaire pour prédire le PnL produisait un R^2 proche de 0 car quelques paires extrêmes dominaient le modèle.

Solution : Basculer vers une approche de **ranking** pondéré, bien moins sensible aux valeurs aberrantes.

7.2 Pondération des actifs

Problème : J'utilisais initialement une pondération simple sans tenir compte de la volatilité, ce qui faussait l'interprétation de la régression linéaire.

Solution : Implémenter la pondération basée sur la volatilité inverse, ajustant le bêta en conséquence.

7.3 Biais de performance passée

Problème : Au début du backtest, des grandes pertes apparaissaient après des grands gains, dégradant le Sharpe ratio.

Solution : Introduire une pénalité pour les paires ayant performé au mois précédent, évitant que le ranking soit biaisé par les succès récents.

8 Pistes d'amélioration

8.1 Intégration de facteurs macroéconomiques

Le pair trading repose actuellement uniquement sur les indicateurs mathématiques. Ajouter une pénalité liée à l'attention macroéconomique (paires où l'une des actions est en focus médiatique) pourrait améliorer la robustesse en évitant les périodes où la cointégration se brise.

8.2 Leviers et stratégies short/long

Nous constatons que 75 % des mois sont positifs, mais avec des gains pas toujours très importants. Utiliser des leviers et des stratégies long/short pourrait augmenter les retours, bien que cela augmente aussi le risque.

8.3 Stop loss et gestion du risque

Pour le moment, nous n'avons pas rencontré de « grandes » pertes mensuelles. Cependant, si une paire sous-performe significativement et représente une part importante du portefeuille, cela pourrait détruire les performances mensuelles. Un **stop loss** bien calibré serait nécessaire pour une implémentation réelle sur les marchés.

9 Conclusion

Ce projet, initialement conçu pour tester la cointégration entre actifs du S&P500, s'est transformé en la construction d'une véritable stratégie de trading quantitative. Bien que résumé, le processus de développement a été long et enrichissant : de l'exploration initiale à la validation sur plusieurs années de données.

Étant mon premier projet quantitatif réel, j'ai appris de nombreuses techniques fondamentales : manipulation de séries temporelles, régression statistique, optimisation de stratégies, et gestion du risque. La familiarisation avec les outils comme *yfinance*, *pandas*, *numpy* et les tests statistiques (Dickey-Fuller, corrélations) me sera certainement utile pour les projets à venir.

Bien que théoriquement prometteuse après calibrage, cette stratégie demanderait des améliorations supplémentaires avant une implémentation réelle sur les marchés financiers. La robustesse des résultats et la faiblesse du drawdown sont encourageantes pour poursuivre ce travail.

Hâte de m'attaquer au prochain défi !